A priori error analysis of an Euler implicit, spectral discretization of Richards equation

N. Abdellatif[a,b], C. Bernardi[c], M. Touihri[d] and D. Yakoubi[e]

[a] University of Manouba, ENSI, Campus Universitaire de Manouba, 2010 Manouba, Tunisie.
[b] University of Tunis El Manar, ENIT, LAMSIN, BP 37, Le Belvédère, 1002 Tunis, Tunisie.
(E-mail: nahla.abdellatif@ensi-uma.tn)

[c] Laboratoire Jacques-Louis Lions, C.N.R.S. & Université Pierre et Marie Curie,
B.C. 187, 4 place Jussieu, 75252 Paris Cedex 05, France.
(E-mail: bernardi@ann.jussieu.fr)

[d] Ecole de l'aviation de Borj El Amri, Tunisia.
(E-mail: moncef_touihri@yahoo.fr)

[e] The Fields Institute for Research in Mathematical Sciences,
222 College Street, Toronto, Ontario M5T 3J1, Canada.
(E-mail: dyakoubi@fields.utoronto.ca)

Abstract

The aim of this work is the numerical study of Richards equation, which models the water flow in a partially saturated underground porous medium under the surface. We propose a discretization of this equation that combines Euler's implicit scheme in time and spectral methods in space. We prove optimal error estimates between the continuous and discrete solutions. Some numerical experiments confirm the interest of this approach. We present numerical experiments which are in perfect coherence with the analysis.

Keywords: Richards equation, spectral methods, space and time discretization, a priori analysis.

1 Introduction

The mass conservation equation:

$$\partial_t \tilde{\Theta}(h_\omega) + \nabla \cdot \mathbf{q}_\omega = 0,$$

which comes in form of volume conservation assuming the incompressibility of the fluid, and Darcy's law:

$$\mathbf{q}_\omega = -K_\omega(\Theta(h_\omega))\nabla(h_\omega + z)$$

lead to the following equation, introduced by Richards in [22]

$$\partial_t \tilde{\Theta}(h_\omega) - \nabla \cdot \big(K_\omega(\Theta(h_\omega))\nabla(h_\omega + z)\big) = 0, \tag{1.1}$$

considered in its pressure formulation. Here, $\mathbf{q}_\omega$ stands for the flux of water and h_ω for the pressure head. The coefficient Θ is the saturation, $\tilde{\Theta}$ is a perturbation of it, K_ω represents

the conductivity and z is the height against the gravitational direction. This equation models the flow of a wetting fluid, mainly water, in the underground surface, hence in an unsaturated medium. In opposite to Darcy's or Brinkman's systems (see [21] for all these models), this equation is nonlinear: Indeed, due to the presence of air above the surface, the porous medium is only partially saturated with water.

The key argument for the analysis of problem (1.1) is to use Kirchhoff's change of unknowns. After this transformation, the new equation fits the general framework proposed in [1] but is simpler. Thus, the existence and uniqueness of a solution to this equation when provided with appropriate initial and boundary conditions are easily derived from standard arguments.

A large number of papers deal with the discretization of similar problems, see e.g. [12] and [13] for finite element discretizations and also [10] for a finite volume one. More recently, several discretizations of Richards equation have been proposed in [4], [11], [20], [25], [27], and [29], see also [28] for a more general equation. All of them rely on a mixed formulation of the previous equation, where the flux $\mathbf{q}_w$ is introduced as a second unknown. We recall this mixed formulation and its well-posedness.

Problem (1.1) is usually discretized in time by Euler's implicit scheme, we also use this scheme for its simplicity. However, it seems that no spectral discretization of this problem has been considered up to now. The aim of this paper is the study of a discretization that combines the Euler's scheme in time and a spectral method in space. We prove the well-posedness of the discrete problem. Next, we establish a priori error estimates that turn out to be fully optimal.

We discuss finally an iterative scheme to solve the nonlinear problems resulting from the discretization procedure. The algorithm we consider was introduced in [26] to construct an effective iteration scheme for fast diffusion problems. It is also used in [19] for mixed finite elements formulation including an implicit discretization of convection. We prove its convergence for the problem that we consider. Some numerical experiments confirm the interest of this approach.

An outline of the paper is as follows.

- In Section 2, we recall the variational formulation of problem (1.1) and its well-posedness. We also write its mixed formulation.
- Section 3 is devoted to the description of the time semi-discrete problem. We recall its well-posedness.
- In Section 4, we write the fully discrete problem and there also check its well-posedness.
- Section 5 is devoted to the a priori error analysis of the discretization.
- Some numerical experiments are presented in Section 6.

2 The continuous problem and its well-posedness

Let Ω be a bounded connected open set in $\mathbb{R}^d$, $d = 2$ or 3, with a Lipschitz-continuous boundary $\partial\Omega$, and let $\mathbf{n}$ denote the unit outward normal vector to Ω on $\partial\Omega$. We assume that $\partial\Omega$ admits a partition without overlap into two parts Γ_D and Γ_F, and that Γ_D has a positive measure. Let also T be a positive real number.

We use here the following Kirchhoff's change of unknowns:

$$x \mapsto \mathcal{K}(x) = \int_0^x K_\omega(\Theta(s))ds.$$

Indeed, by setting

$$u = \mathcal{K}(h_\omega), \quad b(u) = \Theta \circ \mathcal{K}^{-1}(u), \quad k \circ b(u) = K_\omega \circ \Theta \circ \mathcal{K}^{-1}(u)$$

we obtain the following system, where the first line is equivalent to problem (1.1) for a specific choice of $\tilde{\Theta} - \Theta$:

$$\begin{cases} \alpha\, \partial_t u + \partial_t b(u) - \nabla \cdot \Big(\nabla u + k \circ b(u)\mathbf{e}_z\Big) = 0 & \text{in } \Omega\times]0,T[, \\ u = u_D & \text{on } \Gamma_D\times]0,T[, \\ \Big(\nabla u + k \circ b(u)\mathbf{e}_z\Big) \cdot \mathbf{n} = f & \text{on } \Gamma_F\times]0,T[, \\ u|_{t=0} = u_0 & \text{in } \Omega. \end{cases} \tag{2.1}$$

Here $-\mathbf{e}_z$ stands for the unit vector in the direction of gravity. The unknown is now the quantity u. The coefficients b and k are supposed to be known, and their properties are made precise later on, while α is a positive constant. The data are the Dirichlet boundary condition u_D on Γ_D and the initial condition u_0 on Ω, together with the boundary condition f on the normal component of the flux.

In what follows, we use the whole scale of Sobolev spaces $H^m(\Omega)$, with $m \geq 0$, equipped with the norm $\|\cdot\|_{H^m(\Omega)}$ and seminorm $|\cdot|_{H^m\Omega)}$. For any separable Banach space E equipped with the norm $\|\cdot\|_E$, we denote by $\mathscr{C}^0(0,T;E)$ the space of continuous functions from $[0,T]$ with values in E. For each integer $m \geq 0$, we also introduce the space $H^m(0,T;E)$ as the space of measurable functions on $]0,T[$ with values in E such that the mappings: $v \mapsto \|\partial_t^\ell v\|_E$, $0 \leq \ell \leq m$, are square-integrable on $]0,T[$. Finally, we need the spaces $L^\infty(\Omega)$ and $L^\infty(\Omega\times]0,T[)$ of essentially bounded functions on Ω and $\Omega\times]0,T[$, respectively. We are led to make the following assump concerning the coefficients and the data.

Assumption 2.1.
(i) The mapping b is of class $\mathscr{C}^1$, non-decreasing and globally Lipschitz–continuous on $\mathbb{R}$;
(ii) The mapping: $x \mapsto k\circ b(x)$ is continuous, bounded on $\mathbb{R}$ and satisfies for a positive constant c_k

$$\forall x_1 \in \mathbb{R}, \forall x_2 \in \mathbb{R}, \quad \big|k \circ b(x_1) - k \circ b(x_2)\big|^2 \leq c_k \big(b(x_1) - b(x_2)\big)(x_1 - x_2); \tag{2.2}$$

(iii) The function u_0 belongs to $H^1(\Omega)$;
(iv) The function u_D admits a lifting, still denoted by u_D for simplicity, which belongs to $L^2(0,T;H^1(\Omega)) \cap H^1(0,T;L^2(\Omega))$ and satisfies $u_D(\cdot,0) = u_0$;
(v) The function f belongs to $H^1(0,T;L^2(\Gamma_F))$.

In order to take into account the boundary condition on Γ_D, we now introduce the space

$$H^1_D(\Omega) = \big\{ v \in H^1(\Omega);\ v = 0 \text{ on } \Gamma_D \big\}. \tag{2.3}$$

We denote by $H_D^{-1}(\Omega)$ its dual space and by $\langle \cdot, \cdot \rangle$ the duality pairing between $H_D^{-1}(\Omega)$ and $H^1_D(\Omega)$. Next, we consider the following variational problem

Find u in $L^2(0,T;H^1(\Omega))$ with $\partial_t u$ in $L^2(0,T;H_D^{-1}(\Omega))$ such that

$$u = u_D \quad \text{on } \Gamma_D \times]0,T[\qquad \text{and} \qquad u|_{t=0} = u_0 \quad \text{in } \Omega, \tag{2.4}$$

and, for a.e. t in $]0,T[$,

$$\begin{aligned} \forall v \in H^1_D(\Omega), \quad & \alpha \langle \partial_t u(\cdot,t), v \rangle + \langle \partial_t b(u)(\cdot,t), v \rangle \\ & + \int_\Omega \Big(\nabla u + k \circ b(u) \mathbf{e}_z \Big)(\mathbf{x},t) \cdot (\nabla v)(\mathbf{x})\, d\mathbf{x} = \int_{\Gamma_F} f(\tau,t) v(\tau)\, d\tau. \end{aligned} \tag{2.5}$$

From now on, we assume that the partition of $\partial\Omega$ into Γ_D and Γ_F is sufficiently smooth for $\mathscr{D}(\Omega \cup \Gamma_F)$ to be dense into $H^1_D(\Omega)$ (sufficient conditions for this are given in [3] for instance). Then, problem (2.4) – (2.5) is fully equivalent to system (2.1) (in the distribution sense). We refer to [[15],§2.1] for a detailed proof of the next result relying on the monotonicity of the function b.

Theorem 2.2. *If assump* 2.1 *is satisfied, problem* (2.4) – (2.5) *has a unique solution* u. *Moreover, the quantities* $\partial_t u$ *and* $\partial_t b(u)$ *belong to* $L^2(0,T;L^2(\Omega))$.

To go further, we prove a stability property of the solution u exhibited in Theorem 2.2.

Proposition 2.3. *If assump* 2.1 *is satisfied and moreover*
– *the mapping* k *is of class* $\mathscr{C}^1$ *and globally Lipschitz–continuous on* $\mathbb{R}$,
– *the function* u_D *belongs to* $H^1(0,T;H^1(\Omega))$,
the following estimate holds for the solution u *of problem* (2.4) – (2.5), *for all* t *in* $]0,T[$,

$$\begin{aligned} \alpha \int_0^t \|\partial_t u(\cdot,s)\|^2_{L^2(\Omega)}\, ds + |u(\cdot,t)|^2_{H^1(\Omega)} & \\ \leq c \big(1 + \|u_D\|^2_{H^1(0,t;H^1(\Omega))} + \|f\|^2_{H^1(0,t;L^2(\Gamma_F))} \big), & \end{aligned} \tag{2.6}$$

where the constant c *only depends on* α *and* T.

Proof: We set:

$$u(\mathbf{x},t) = u_D(\mathbf{x},t) + u_*(\mathbf{x},t), \qquad b_*(w) = b(u_D + w).$$

Thus, it is readily checked that u_* belongs to $L^2(0,T;H^1_D(\Omega))$ and satisfies

$$\begin{aligned} \forall v \in H^1_D(\Omega), \quad & \alpha \langle \partial_t u_*(\cdot,t), v \rangle + \langle \partial_t b_*(u_*)(\cdot,t), v \rangle \\ & + \int_\Omega \Big(\nabla u_* + \nabla u_D + k \circ b_*(u_*) \mathbf{e}_z \Big)(\mathbf{x},t) \cdot (\nabla v)(\mathbf{x})\, d\mathbf{x} \\ & = -\alpha \langle \partial_t u_D(\cdot,t), v \rangle + \int_{\Gamma_F} f(\tau,t) v(\tau)\, d\tau, \end{aligned}$$

Next, we formally take v equal to $\partial_t u_*(\cdot,t)$ (i.e., we use a regularization of it when necessary) and integrate the equation with respect to t. Since b_*' is nonnegative, this leads to (note that

u_* vanishes at $t=0$)

$$\alpha\int_0^t \|\partial_t u_*(\cdot,s)\|^2_{L^2(\Omega)}\,ds+\frac12|u_*(\cdot,t)|^2_{H^1(\Omega)}\le -\int_0^t\int_\Omega(\nabla u_D)(\mathbf{x},s)\,\cdot\,(\nabla\partial_t u_*)(\mathbf{x},s)\,d\mathbf{x}ds$$
$$-\int_0^t\int_\Omega\big(k\circ b_*(u_*)\mathbf{e}_z\big)(\mathbf{x},s)\,\cdot\,(\nabla\partial_t u_*)(\mathbf{x},s)\,d\mathbf{x}ds$$
$$-\alpha\int_0^t\langle\partial_t u_D(\cdot,s),\partial_t u_*(\cdot,s)\rangle\,ds+\int_0^t\int_{\Gamma_F} f(\tau,s)\partial_t u_*(\tau,s)\,d\tau ds.$$

To handle the third integral in the right-hand side, we use Young's inequality. To handle the other ones, we integrate by parts with respect to t. All this yields

$$\frac{\alpha}{2}\int_0^t \|\partial_t u_*(\cdot,s)\|^2_{L^2(\Omega)}\,ds+\frac12|u_*(\cdot,t)|^2_{H^1(\Omega)}$$
$$\le|u_D(\cdot,t)|_{H^1(\Omega)}|u_*(\cdot,t)|_{H^1(\Omega)}+\int_0^t\int_\Omega(\nabla\partial_t u_D)(\mathbf{x},s)\,\cdot\,(\nabla u_*)(\mathbf{x},s)\,d\mathbf{x}ds$$
$$+\int_0^t\int_\Omega\big(k\circ b_*\big)'(u_*)(\mathbf{x},s)(\partial_t u_D+\partial_t u_*)(\mathbf{x},s)\mathbf{e}_z\,\cdot\,(\nabla u_*)(\mathbf{x},s)\,d\mathbf{x}ds$$
$$+c\,|u_*(\cdot,t)|_{H^1(\Omega)}+\frac{\alpha}{2}\|\partial_t u_D\|^2_{L^2(0,t;L^2(\Omega))}$$
$$+c\,\|f(\cdot,t)\|_{L^2(\Gamma_F)}\,|u_*(\cdot,t)|_{H^1(\Omega)}-\int_0^t\int_{\Gamma_F}\partial_t f(\tau,s)u_*(\tau,s)\,d\tau ds.$$

Since k' and b' are bounded, we conclude by using appropriate Young's inequalities and the Grönwall's lemma.

In view of the discretization, we finally introduce a mixed formulation of problem (2.4) – (2.5). To this aim, we consider the domain $H(\mathrm{div},\Omega)$ of the divergence operator, namely

$$H(\mathrm{div},\Omega)=\big\{\varphi\in L^2(\Omega)^d;\ \nabla\,\cdot\,\varphi\in L^2(\Omega)\big\},\tag{2.7}$$

equipped with the graph norm. Since the normal trace operator: $\varphi\mapsto\varphi\cdot\mathbf{n}$ can be defined from $H(\mathrm{div},\Omega)$ onto $H^{-\frac12}(\partial\Omega)$, see e.g. [[14], Chap. I, Thm 2.5], and its restriction to Γ_F maps $H(\mathrm{div},\Omega)$ into the dual space of $H^{\frac12}_{00}(\Gamma_F)$ (see [[**?**], Chap. 1, Th. 11.7] for the definition of this last space), we also introduce the space

$$H_F(\mathrm{div},\Omega)=\big\{\varphi\in H(\mathrm{div},\Omega);\ \varphi\cdot\mathbf{n}=0\text{ on }\Gamma_F\big\}.\tag{2.8}$$

The mixed variational problem then reads

Find $(u,\mathbf{q})$ in $L^2(0,T;L^2(\Omega))\times L^2(0,T;H(\mathrm{div},\Omega))$ with $\partial_t u$ in $L^2(0,T;L^2(\Omega))$ such that

$$\mathbf{q}\cdot\mathbf{n}=-f\quad\text{on }\Gamma_F\times]0,T[\qquad\text{and}\qquad u|_{t=0}=u_0\quad\text{in }\Omega,\tag{2.9}$$

and, for a.e. t in $]0,T[$,

$$\begin{aligned}
&\forall w\in L^2(\Omega),\quad \alpha\int_\Omega\big(\partial_t u\big)(\mathbf{x},t)w(\mathbf{x})\,d\mathbf{x}+\int_\Omega\big(\partial_t b(u)\big)(\mathbf{x},t)w(\mathbf{x})\,d\mathbf{x}\\
&\qquad\qquad+\int_\Omega\big(\nabla\,\cdot\,\mathbf{q}\big)(\mathbf{x},t)w(\mathbf{x})\,d\mathbf{x}=0,\\
&\forall\varphi\in H_F(\mathrm{div},\Omega),\quad \int_\Omega\mathbf{q}(\mathbf{x},t)\,\cdot\,\varphi(\mathbf{x})\,d\mathbf{x}-\int_\Omega u(\mathbf{x},t)(\nabla\,\cdot\,\varphi)(\mathbf{x})\,d\mathbf{x}\\
&\qquad\qquad+\int_\Omega\big(k\circ b(u)\big)(\mathbf{x},t)\mathbf{e}_z\,\cdot\,\varphi(\mathbf{x})\,d\mathbf{x}=-\langle u_D(\cdot,t),\varphi\cdot\mathbf{n}\rangle_{\Gamma_D},
\end{aligned}\tag{2.10}$$

where $\langle\cdot,\cdot\rangle_{\Gamma_D}$ now denotes the duality pairing between $H^{\frac{1}{2}}(\Gamma_D)$ and its dual space. The following equivalence property is readily checked, see [[4], Prop. 2.6].

Proposition 2.4. *If assump* 2.1 *is satisfied, problems* (2.4) − (2.5) *and* (2.9) − (2.10) *are equivalent, in the following sense:*
(i) For any solution u *of* $(2.4)-(2.5)$*, there exists a function* $\mathbf{q}$ *in* $L^2(0,T;H(\mathrm{div},\Omega))$ *such that the pair* $(u,\mathbf{q})$ *is a solution of problem* $(2.9)-(2.10)$*;*
(ii) For any solution $(u,\mathbf{q})$ *of* $(2.9)-(2.10)$*, the function* u *belongs to* $L^2(0,T;H^1(\Omega))$ *and is a solution of problem* (2.4) − (2.5)*.*

As a direct consequence of Theorem 2.2 and Proposition 2.4, if assump 2.1 is satisfied, problem $(2.9)-(2.10)$ has a unique solution $(u,\mathbf{q})$. Note moreover that, in contrast with u, the flux $\mathbf{q}$ has a physical meaning. Indeed, it is equal to

$$\mathbf{q} = -\nabla u - k\circ b(u)\mathbf{e}_z, \tag{2.11}$$

and thus coincides with the flux $\mathbf{q}_\omega$ in problem (1.1).

3 The time semi-discrete problem and its well-posedness

We now propose a time semi-discretization relying on the mixed formulation $(2.9)-(2.10)$. The next analysis requires hypotheses which are slightly stronger than Assumption 2.1 but still not restrictive.

Assumption 3.1.
(i) The mappings b *and* k *and the data* u_0*,* u_D*, and* f *satisfy Assumption 2.1;*
(ii) The function k *is Lipschitz–continuous on* $\mathbb{R}$*;*
(iii) The function u_D *belongs to* $\mathcal{C}^0(0,T;H^1(\Omega))$*.*

Since we intend to work with non uniform time steps, we introduce a partition of the interval $[0,T]$ into subintervals $[t_{j-1},t_j]$, $1\le j\le J$, such that $0=t_0<t_1<\cdots<t_J=T$. We denote by τ_j the time step t_j-t_{j-1}, by τ the J-tuple $(\tau_1,\ldots,\tau_J)$ and by $|\tau|$ the maximum of the τ_j, $1\le j\le J$.

As already hinted in Section 1, the time discretization mainly relies on a backward Euler's scheme, where the nonlinear term $k\circ b(u)$ is treated in an explicit way for simplicity. Thus, the semi-discrete problem reads:

Find $(u^j)_{0\le j\le J}$ in $L^2(\Omega)^{J+1}$ and $(\mathbf{q}^j)_{1\le j\le J}$ in $H(\mathrm{div},\Omega)^J$ such that

$$\mathbf{q}^j\cdot\mathbf{n} = -f(\cdot,t_j)\quad\text{on }\Gamma_F,\ 1\le j\le J,\qquad\text{and}\qquad u^0=u_0\quad\text{in }\Omega,$$

and, for $1\le j\le J$,

$$\begin{aligned}
&\forall w\in L^2(\Omega),\\
&\qquad\alpha\int_\Omega\Big(\frac{u^j-u^{j-1}}{\tau_j}\Big)(\mathbf{x})w(\mathbf{x})\,d\mathbf{x}+\int_\Omega\Big(\frac{b(u^j)-b(u^{j-1})}{\tau_j}\Big)(\mathbf{x})w(\mathbf{x})\,d\mathbf{x}\\
&\qquad\qquad+\int_\Omega\big(\nabla\cdot\mathbf{q}^j\big)(\mathbf{x})w(\mathbf{x})\,d\mathbf{x}=0,\\
&\forall\varphi\in H_F(\mathrm{div},\Omega),\quad\int_\Omega\mathbf{q}^j(\mathbf{x})\cdot\varphi(\mathbf{x})\,d\mathbf{x}-\int_\Omega u^j(\mathbf{x})(\nabla\cdot\varphi)(\mathbf{x})\,d\mathbf{x}\\
&\qquad\qquad+\int_\Omega\big(k\circ b(u^{j-1})\big)(\mathbf{x})\mathbf{e}_z\cdot\varphi(\mathbf{x})\,d\mathbf{x}=-\langle u_D(\cdot,t_j),\varphi\cdot\mathbf{n}\rangle_{\Gamma_D}.
\end{aligned}$$

It can be noted that this problem makes sense since both u_D and f are continuous in time. We recall its well-posedness from [[4], Prop. 3.2] (note that this still requires the density of $\mathcal{D}(\Omega \cup \Gamma_F)$ in $H^1_D(\Omega)$).

Proposition 3.2. *If Assumption* 3.1 *is satisfied, problem* (3.1) − (3.2) *has a unique solution* $(u^j, \mathbf{q}^j)_j$.

In analogy with Proposition 2.3, we prove a stability property of the solution $(u^j, \mathbf{q}^j)$ which is needed later on.

Lemma 3.3. *If Assumption* 3.1 *is satisfied, the following estimate holds for the solutions* $(u^j, \mathbf{q}^j)$ *of problems* (3.1) − (3.2), $1 \le j \le J$,

$$\begin{aligned}
&\Big(\alpha \sum_{m=1}^{j} \tau_m \, \|\frac{u^m - u^{m-1}}{\tau_m}\|^2_{L^2(\Omega)}\Big)^{\frac{1}{2}} + |u^j|_{H^1(\Omega)} \\
&\qquad \le c \Bigg(\sqrt{j} + |u_0|_{H^1(\Omega)} \\
&\qquad\quad + \Big(\sum_{m=1}^{j} \tau_m \, \|\frac{u_D(\cdot,t_m) - u_D(\cdot,t_{m-1})}{\tau_m}\|^2_{L^2(\Omega)}\Big)^{\frac{1}{2}} + \Big(\sum_{m=1}^{j} |u_D(\cdot,t_m)|^2_{H^1(\Omega)}\Big)^{\frac{1}{2}} \\
&\qquad\qquad\qquad + \Big(\sum_{m=1}^{j} \|f(\cdot,t_m)\|^2_{L^2(\Gamma_F)}\Big)^{\frac{1}{2}} \Bigg).
\end{aligned}$$

Proof: Setting as previously $u^j = u^j_* + u_D(\cdot,t_j)$, we observe that problem (3.2) can equivalent be written as

$$\begin{aligned}
\forall v \in H^1_D(\Omega), \quad & \alpha \, \langle \frac{u^j_* - u^{j-1}_*}{\tau_j}, v\rangle + \langle \frac{b(u^j_* + u_D(\cdot,t_j)) - b(u^{j-1}_* + u_D(\cdot,t_j))}{\tau_j}, v\rangle \\
& \qquad + \int_\Omega \big(\nabla u^j_* + k \circ b(u^{j-1}) \mathbf{e}_z\big)(\mathbf{x}) \cdot (\nabla v)(\mathbf{x}) \, d\mathbf{x} = \langle g^j, v\rangle + \langle h^j, v\rangle,
\end{aligned}$$

where the quantities g^j and h^j are defined by

$$\begin{aligned}
\langle g^j, v\rangle &= -\alpha \, \langle \frac{u_D(\cdot,t_j) - u_D(\cdot,t_{j-1})}{\tau_j}, v\rangle \\
&\qquad - \langle \frac{b\big(u^{j-1}_* + u_D(\cdot,t_j)\big) - b\big(u^{j-1}_* + u_D(\cdot,t_{j-1})\big)}{\tau_j}, v\rangle \\
\langle h^j, v\rangle &= -\int_\Omega (\nabla u_D)(\mathbf{x},t_j) \cdot (\nabla v)(\mathbf{x}) \, d\mathbf{x} + \int_{\Gamma_F} f(\tau,t_j) v(\tau) \, d\tau.
\end{aligned}$$

Thus, taking v equal to $u^j_* - u^{j-1}_*$, noting that the quantity

$$\langle b(u^j_* + u_D(\cdot,t_j)) - b(u^{j-1}_* + u_D(\cdot,t_j)), u^j_* - u^{j-1}_*\rangle$$

is nonnegative and using the formula

$$\nabla u^j_* \cdot \nabla(u^j_* - u^{j-1}_*) = \frac{1}{2}\Big(|\nabla(u^j_* - u^{j-1}_*)|^2 + |\nabla u^j_*|^2 - |\nabla u^{j-1}_*|^2\Big),$$

together with the boundedness of the mapping k lead to

$$\begin{aligned}
&\alpha \, \tau_j \, \|\frac{u^j_* - u^{j-1}_*}{\tau_j}\|^2_{L^2(\Omega)} + \frac{1}{2}|u^j_*|^2_{H^1(\Omega)} + \frac{1}{2}|u^j_* - u^{j-1}_*|^2_{H^1(\Omega)} \\
&\le \frac{1}{2}|u^{j-1}_*|^2_{H^1(\Omega)} + (c + \|h^j\|_{H^{-1}_D(\Omega)})|u^j_* - u^{j-1}_*|_{H^1(\Omega)} + \tau_j \, \|g^j\|_{L^2(\Omega)} \|\frac{u^j_* - u^{j-1}_*}{\tau_j}\|_{L^2(\Omega)}.
\end{aligned}$$

By using the inequality $ab \leq \frac{1}{2}(a^2+b^2)$ and summing on the j, we obtain

$$\frac{\alpha}{2}\sum_{m=1}^{j}\tau_m\,\|\frac{u_*^m-u_*^{m-1}}{\tau_m}\|^2_{L^2(\Omega)}+\frac{1}{2}|u_*^j|^2_{H^1(\Omega)} \\ \leq cj+\sum_{m=1}^{j}\|h^m\|^2_{H_D^{-1}(\Omega)}+\frac{1}{2\alpha}\sum_{m=1}^{j}\tau_m\,\|g^m\|^2_{L^2_D(\Omega)}.$$

We conclude thanks to the definition of g^j and h^j and by using a triangle inequality.

Remark 3.4. From now on, we denote by $c_0(\tau)$ the maximum of the quantities that appear in the right-hand side of estimate (3.3), namely

$$c_0(\tau)=c\Big(\sqrt{J}+|u_0|_{H^1(\Omega)} \\ +\Big(\sum_{m=1}^{J}\tau_m\,\|\frac{u_D(\cdot,t_m)-u_D(\cdot,t_{m-1})}{\tau_m}\|^2_{L^2(\Omega)}\Big)^{\frac{1}{2}}+\Big(\sum_{m=1}^{J}|u_D(\cdot,t_m)|^2_{H^1(\Omega)}\Big)^{\frac{1}{2}} \\ +\Big(\sum_{m=1}^{J}\|f(\cdot,t_m)\|^2_{L^2(\Gamma_F)}\Big)^{\frac{1}{2}}\Big). \tag{3.1}$$

There is no reason for the last terms in this quantity to be bounded independently of τ. Assumption 3.1 only implies that

$$c_0(\tau)\leq c\sqrt{J}. \tag{3.2}$$

4 The time and space discrete problem and its well-posedness

We assume that Ω is the square or cube $]-1,1[^d$, $d=2$ or 3, and we suppose that Γ_D is a union of whole edges ($d=2$) or faces ($d=3$) of Ω. In order to describe the discrete spaces and for any triple (ℓ,m,n) of nonnegative integers, we introduce
• in dimension $d=2$, the space $\mathbb{P}_{\ell,m}(\Omega)$ of restrictions to Ω of polynomials with degree $\leq\ell$ with respect to x and $\leq m$ with respect to y,
• in dimension $d=3$, the space $\mathbb{P}_{\ell,m,n}(\Omega)$ of restrictions to Ω of polynomials with degree $\leq\ell$ with respect to x, $\leq m$ with respect to y and $\leq n$ with respect to z.
When ℓ and m are equal to n, these spaces are denoted by $\mathbb{P}_n(\Omega)$.

We fix an integer $N\geq 2$ and we approximate $L^2(\Omega)$ by

$$\mathbb{X}_N=\mathbb{P}_{N-1}(\Omega). \tag{4.1}$$

In analogy with [[18], Def. 7], the space which approximates $H(\mathrm{div},\Omega)$ is defined by

$$\mathbb{Y}_N=\begin{cases}\mathbb{P}_{N,N-1}(\Omega)\times\mathbb{P}_{N-1,N}(\Omega) & \text{if } d=2,\\ \mathbb{P}_{N,N-1,N-1}(\Omega)\times\mathbb{P}_{N-1,N,N-1}(\Omega)\times\mathbb{P}_{N-1,N-1,N}(\Omega) & \text{if } d=3.\end{cases} \tag{4.2}$$

We also introduce the space

$$\mathbb{Y}_{NF}=\mathbb{Y}_N\cap H_F(\mathrm{div},\Omega). \tag{4.3}$$

In order to handle nonregular functions b and k, as first suggested in [17], we possibly use over-integration. We fix an integer $M\geq N$ and, setting $\xi_{M0}=-1$ and $\xi_{MM}=1$, we introduce

the $M-1$ nodes ξ_{Mi}, $1 \le i \le M-1$, and the $M+1$ weights ρ_{Mi}, $0 \le i \le M$, of the Gauss-Lobatto quadrature formula:

$$\forall\ \Phi\ \in\ \mathbb{P}_{2M-1}(-1,1), \quad \int_{-1}^{1} \Phi(x)dx = \sum_{i=0}^{M} \Phi(\xi_{Mi})\rho_{Mi} \tag{4.4}$$

with obvious definition for the spaces $\mathbb{P}_n(-1,1)$. We also recall from [[5], Eq. (13.20)]the basic property of this formula, which is needed in what follows:

$$\forall \varphi_M \in \mathbb{P}_M(-1,1), \quad \|\varphi_M\|^2_{L^2(-1,1)} \le \sum_{i=0}^{M} \varphi_M^2(\xi_{Mi})\rho_{Mi} \le 3\,\|\varphi_M\|^2_{L^2(-1,1)}. \tag{4.5}$$

Relying on this formula, we define the discrete product, for continuous functions φ and ψ by

$$(\varphi,\psi)_M = \begin{cases} \displaystyle\sum_{i=0}^{M}\sum_{j=0}^{M} \varphi(\xi_{Mi},\xi_{Mj})\psi(\xi_{Mi},\xi_{Mj})\rho_{Mi}\rho_{Mj} & \text{if } d=2, \\ \displaystyle\sum_{i=0}^{M}\sum_{j=0}^{M}\sum_{k=0}^{M} \varphi(\xi_{Mi},\xi_{Mj},\xi_{Mk})\psi(\xi_{Mi},\xi_{Mj},\xi_{Mk})\rho_{Mi}\rho_{Mj}\rho_{Mk} & \text{if } d=3. \end{cases}$$

Similarly, we define a discrete product $(\cdot,\cdot)_M^{\Gamma_D}$ on Γ_D (recall that it is a union of whole edges or faces of Ω). Let $\mathcal{I}_M$ denote the Lagrange interpolation operator at the nodes of the grid

$$\Sigma_M = \begin{cases} \big\{(\xi_{Mi},\xi_{Mj}),\ 0 \le i,j \le M\big\} & \text{if } d=2, \\ \big\{(\xi_{Mi},\xi_{Mj},\xi_{Mk}),\ 0 \le i,j,k \le M\big\} & \text{if } d=3, \end{cases}$$

with values in $\mathbb{P}_M(\Omega)$. Finally, let i_M^{Γ} stand for the Lagrange interpolation operator at the nodes of $\Sigma_M \cap \Gamma$, where Γ is either Γ_F or Γ_D, with values in the space of traces of $\mathbb{P}_M(\Omega)$ onto Γ. The fully discrete problem reads:

Find $(u_N^j)_{0\le j\le J}$ in $\mathbb{X}_N^{J+1}$ and $(\mathbf{q}_N^j)_{1\le j\le J}$ in $\mathbb{Y}_N^J$ such that

$$\mathbf{q}_N^j \cdot \mathbf{n} = -i_{N-1}^{\Gamma_F} f(\cdot,t_j) \quad \text{on } \Gamma_F,\ 1 \le j \le J, \qquad \text{and} \qquad u_N^0 = \mathcal{I}_{N-1}u_0 \quad \text{in } \Omega, \tag{4.6}$$

and, for $1 \le j \le J$,

$$\begin{aligned} &\forall w_N \in \mathbb{X}_N, \quad \alpha\Big(\frac{u_N^j - u_N^{j-1}}{\tau_j}, w_N\Big)_M + \Big(\frac{b(u_N^j) - b(u_N^{j-1})}{\tau_j}, w_N\Big)_M + (\nabla\cdot\mathbf{q}_N^j, w_N)_M = 0, \\ &\forall \varphi_N \in \mathbb{Y}_{NF}, \quad (\mathbf{q}_N^j, \varphi_N)_M - (u_N^j, \nabla\cdot\varphi_N)_M + (\mathcal{I}_{N-1}(k\circ b(u_N^{j-1})\mathbf{e}_z), \varphi_N)_M \\ &\qquad\qquad = -(u_{DN}^j, \varphi_N\cdot\mathbf{n})_M^{\Gamma_D}, \end{aligned} \tag{4.7}$$

where we denote by u_{DN}^j the function $i_N^{\Gamma_D}u_D(.,t_j)$, $1 \le j \le J$. The statement of this problem requires some further pointwise continuity of the data, so that we are led to make the following assumption.

Assumption 4.1.
(i) The mappings b and k and the data u_0, u_D, and f satisfy Assumption 3.1;
(ii) The function u_0 belongs to $H^{s_1}(\Omega)$, $s_1 > \frac{d}{2}$;
(iii) The function u_D belongs to $\mathcal{C}^0(0,T;H^{s_2}(\Gamma_D))$, $s_2 > \frac{d-1}{2}$;
(iv) The function f belongs to $\mathcal{C}^0(0,T;H^{s_3}(\Gamma_F))$, $s_3 > \frac{d-1}{2}$.

We now intend to prove the following statement.

Proposition 4.2. *If Assumption* 4.1 *is satisfied, problem* $(4.6)-(4.7)$, $1 \leq j \leq J$, *has a unique solution* $(u_N^j, \mathbf{q}_N^j)_j$.

Proof: We proceed by induction on j and we check successively the uniqueness of the solution, next its existence.
1) Let $(u_N^j, \mathbf{q}_N^j)$ and $(\tilde{u}_N^j, \tilde{\mathbf{q}}_N^j)$ be two solutions of problem (4.6)-(4.7). Then the pair $(u_N^j - \tilde{u}_N^j, \mathbf{q}_N^j - \tilde{\mathbf{q}}_N^j)$ satisfies:

$$\begin{aligned}
&\forall w_N \in X_N, \\
&\qquad \alpha(\frac{u_N^j - \tilde{u}_N^j}{\tau_j}, w_N)_M + (\frac{b(u_N^j) - b(\tilde{u}_N^j)}{\tau_j}, w_N)_M + (\nabla \cdot (q_N^j - \tilde{\mathbf{q}}_N^j), w_N)_M = 0, \qquad (4.8)\\
&\forall \varphi_N \in Y_{NF}, \quad (q_N^j - \tilde{q}_N^j, \varphi_N)_M - (u_N^j - \tilde{u}_N^j, \nabla \cdot \varphi_N)_M = 0.
\end{aligned}$$

Then, we take $w_N = u_N^j - \tilde{u}_N^j$ and $\varphi_N = \mathbf{q}_N^j - \tilde{\mathbf{q}}_N^j$. We sum the two equations and we use the exactness of the quadrature formula and property (4.5). This yields

$$\frac{\alpha}{\tau_j}\|u_N^j - \tilde{u}_N^j\|^2_{L^2(\Omega)} + \|\mathbf{q}_N^j - \tilde{\mathbf{q}}_N^j\|^2_{L^2(\Omega)^d} \leq -\frac{1}{\tau_j}(b(u_N^j) - b(\tilde{u}_N^j), u_N^j - \tilde{u}_N^j)_M.$$

Since, in dimension $d = 2$ for instance,

$$\begin{aligned}
&(b(u_N^j) - b(\tilde{u}_N^j), u_N^j - \tilde{u}_N^j)_M \\
&= \sum_{i,j=0}^{M} \Big(b(u_N^j(\xi_{Mi}, \xi_{Mj})) - b(\tilde{u}_N^j(\xi_{Mi}, \xi_{Mj}))\Big)\Big(u_N^j(\xi_{Mi}, \xi_{Mj}) - \tilde{u}_N^(\xi_{Mi}, \xi_{Mj})\Big)\rho_{Mi}\rho_{Mj},
\end{aligned}$$

and since the ρ_{Mi} are positive, we use Assumption 4.1 to derive:

$$\frac{\alpha}{\tau_j}\|u_N^j - \tilde{u}_N^j\|^2_{L^2(\Omega)} + \|\mathbf{q}_N^j - \tilde{\mathbf{q}}_N^j\|^2_{L^2(\Omega)^d} \leq 0.$$

So, $(u_N^j, \mathbf{q}_N^j)$ is equal to $(\tilde{u}_N^j, \tilde{\mathbf{q}}_N^j)$, for $1 \leq j \leq J$.
2) Proving the existence of a solution relies on Brouwer's fixed point theorem. Let $\bar{f}(\cdot, t_j)$ be an extension of $f(\cdot, t_j)$ in $L^2(\partial\Omega)$ by 0. From [[14], Chap. I, Corollary 2.8], there exists $\chi^j \in H(\mathrm{div}, \Omega)$ such that $\chi^j \cdot \mathbf{n} = -i_{N-1}^{\partial\Omega}\bar{f}(\cdot, t_j)$, and

$$\|\chi^j\|_{H(\mathrm{div},\Omega)} \leq c_1 \|i_{N-1}^{\partial\Omega}\bar{f}\|_{L^2(\partial\Omega)}. \qquad (4.9)$$

Thus, the function $\chi_N^j = \mathcal{I}_{N-1}\chi^j$ belongs to $\mathbb{Y}_N$ and satisfies, thanks to an appropriate inverse inequality,

$$\|\chi_N^j\|_{H(\mathrm{div},\Omega)} \leq c_2\, N^2 \|f(\cdot, t_j)\|_{H^{s_3}(\Gamma_F)}. \qquad (4.10)$$

Next, we define the mapping Ψ on $\mathbb{X}_N \times \mathbb{Y}_{NF}$ by, for any $(w_N, \varphi_N) \in \mathbb{X}_N \times \mathbb{Y}_{NF}$,

$$\begin{aligned}
< \Psi(u_N, \mathbf{q}_N), (w_N, \varphi_N) > = \alpha(\frac{u_N}{\tau_j}, w_N)_M + (\frac{b(u_N)}{\tau_j}, w_N)_M + (\nabla \cdot \mathbf{q}_N, \varphi_N)_M \\
+(\mathbf{q}_N, \varphi_N)_M - (u_N, \nabla \cdot \varphi_N)_M - R^j(w_N, \varphi_N)
\end{aligned}$$

where

$$R^j(w_N,\varphi_N) = \alpha(\frac{u_N^{j-1}}{\tau_j},w_N)_M + (\frac{b(u_N^{j-1})}{\tau_j},w_N)_M - (\nabla.\chi_N^j,w_N)_M - (\chi_N^j,\varphi_N)_M \\ -(\mathcal{I}_N k\circ b(u_N^{j-1})\mathbf{e}_z,\varphi_N)_M - (i_N^{\Gamma_D}u_D(.,t_j),\varphi_N\cdot\mathbf{n})_M.$$

When providing $\mathbb{X}_N\times\mathbb{Y}_N$ with the norm of $L^2(\Omega)\times H(\mathrm{div},\Omega)$, the mapping Ψ is continuous. Since all norms on this finite-dimensional space are equivalent, it is also continuous when this space is equipped with the norm of $L^2(\Omega)\times L^2(\Omega)^d$. Moreover it satisfies

$$<\Psi(u_N,\mathbf{q}_N),(u_N,\mathbf{q}_N)>=\tfrac{\alpha}{\tau_j}\|u_N\|^2_{L^2(\Omega)}+\tfrac{1}{\tau_j}(b(u_N),u_N)_M+\|\mathbf{q}_N\|^2_{L^2(\Omega)}-R^j(u_N,\mathbf{q}_N),$$

whence

$$<\Psi(u_N,\mathbf{q}_N),(u_N,\mathbf{q}_N)>\geq\frac{\alpha}{\tau_j}\|u_N\|^2_{L^2(\Omega)}+\|\mathbf{q}_N\|^2_{L^2(\Omega)^d}-C_j(\frac{\alpha}{\tau_j}\|u_N\|^2_{L^2(\Omega)}+\|\mathbf{q}_N\|^2_{L^2(\Omega)^d})^{\frac12},$$

where C_j is given by

$$C_j=\tfrac{\alpha}{\tau_j}\|u_N^{j-1}\|_{L^2(\Omega)}+\tfrac{1}{\alpha\tau_j}\|b(u_N^{j-1})\|_{L^2(\Omega)}+\tfrac{c_2\,N^2(1+\alpha)}{\alpha}\|f\|_{H^{s_3}(\Gamma_F)} \\ +\|k\circ b(u_N^{j-1})e_z\|_{L^2(\Omega)}+c\,N^2\,\|u_D(\cdot,t_j)\|_{H^{s_2}(\Gamma_D)}$$

(here also, we made use of the inverse inequality $\|\mathbf{q}_N\|_{H(\mathrm{div},\Omega)}\leq c\,N^2\,\|\mathbf{q}_N\|_{L^2(\Omega)^d}$). It thus follows from Brouwer's fixed point theorem, see [[14], Chap. I, Corollary 1.1] for instance, that there exists a pair (u_N^j,q_{N*}^j) such that

$$\forall(w_N,\varphi_N)\in\mathbb{X}_N\times\mathbb{Y}_{NF},\quad <\Psi(u_N^j,\mathbf{q}_{N*}^j),(w_N,\varphi_N)>=0$$

and

$$\|u_N^j\|_{L^2(\Omega)}+\|q_{N*}^j\|_{L^2\Omega)}\leq C_j.$$

Then, the pair $(u_N^j,\mathbf{q}_N^j=\mathbf{q}_{N*}^j+\chi_N^j)$ is a solution of problem (4.6) – (4.7).

The previous proof does not lead to any appropriate bound for the solution $(u_N^j,\mathbf{q}_N^j)$. However, we do not need it in what follows.

5 A priori error analysis

We evaluate simultaneously the errors due to the time and space discetizations. Indeed, even if they require different regularity properties of the solution $(u,\mathbf{q})$, proving them relies on very similar arguments, more precisely on the theorem of Brezzi, Rappaz and Raviart, see [8]. Only for simplicity, all this analysis is performed in the case $f=0$ of zero Neumann conditions.

5.1 Some preliminary results

Let $\mathcal{T}$ denote the operator associated with the linear heat equation: For any data F in $L^2(0,T;L^2(\Omega))$ and (u_0,u_D) satisfying Assumption 2.1, $\mathcal{T}(F,u_0,u_D)$ is equal to the solution u of the problem

Find u in $\mathscr{C}^0(0,T;H^1(\Omega))\cap H^1(0,T;L^2(\Omega))$ such that

$$u=u_D\quad\text{on }\Gamma_D\times]0,T[\qquad\text{and}\qquad u|_{t=0}=u_0\quad\text{in }\Omega, \tag{5.1}$$

and, for a.e. t in $]0,T[$,

$$\forall v \in H^1_D(\Omega), \quad \alpha \int_\Omega \partial_t u(\mathbf{x},t)v(x)\,d\mathbf{x} + \int_\Omega \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x})\,d\mathbf{x} = \int_\Omega F(\mathbf{x},t)v(\mathbf{x})\,d\mathbf{x}. \tag{5.2}$$

The well-posedness of this problem is obvious.

With this notation, problem (2.4) – (2.5) can equivalently be written

$$\mathcal{F}(u) = u - \mathcal{T}\big(\mathcal{G}(u), u_0, u_D\big) = 0, \tag{5.3}$$

where the mapping $\mathcal{G}$ is defined by

$$\int_\Omega \mathcal{G}(u)(\mathbf{x},t)v(x,t)\,d\mathbf{x} = -\langle \partial_t b(u)(\cdot,t), v\rangle - \int_\Omega k\circ b(u)(\mathbf{x},t)\mathbf{e}_z \cdot \nabla v(\mathbf{x},t)\,d\mathbf{x} \tag{5.4}$$

We first prove a further property of $D\mathcal{F}(u)$, where u is the solution of problem (2.4) – (2.5) and D stands for the differential operator with respect to u. From now on, we set:

$$\mathbb{W} = L^\infty(0,T;H^1(\Omega)) \cap H^1(0,T;L^2(\Omega)). \tag{5.5}$$

Lemma 5.1. *Assume that the mappings b and k are of class $\mathscr{C}^2$ with bounded derivatives. If the solution u of problem (2.4) – (2.5) is such that $\partial_t u$ belongs to $L^\infty(\Omega\times]0,T[)$, the operator $D\mathcal{F}(u)$ is an isomorphism of $\mathbb{W}$.*

Proof: The assertion of the lemma is equivalent to the well-posedness of the problem, for any data F in $L^2(0,T;L^2(\Omega))$:

Find w in $\mathscr{C}^0(0,T;H^1(\Omega)) \cap H^1(0,T;L^2(\Omega))$ such that

$$w = 0 \quad \text{on } \Gamma_D\times]0,T[\qquad \text{and} \qquad w|_{t=0} = 0 \quad \text{in } \Omega, \tag{5.6}$$

and, for a.e. t in $]0,T[$,

$$\begin{aligned}\forall v \in H^1_D(\Omega), \quad & \alpha \int_\Omega \partial_t w(\mathbf{x},t)v(\mathbf{x},t)\,d\mathbf{x} + \int_\Omega \partial_t\big(b'(u)w\big)(\mathbf{x},t)v(\mathbf{x},t)\,d\mathbf{x} \\ & + \int_\Omega \big(\nabla w + (k\circ b)'(u)w\big)(\mathbf{x},t) \cdot \nabla v(\mathbf{x},t)\,d\mathbf{x} = \int_\Omega F(\mathbf{x},t)v(\mathbf{x},t)\,d\mathbf{x}.\end{aligned} \tag{5.7}$$

We proceed in several steps, beginning with an a priori estimate of any solution of this problem.
1) First, we take v equal to w in (5.7) and integrate with respect to t. By noting that

$$\begin{aligned}\int_0^t \int_\Omega \partial_t\big(b'(u)w\big)(\mathbf{x},s)w(\mathbf{x},s)\,d\mathbf{x}ds &= \int_0^t \int_\Omega \big(b''(u)(\partial_t u)w + b'(u)(\partial_t w)\big)(\mathbf{x},s)w(\mathbf{x},s)\,d\mathbf{x} \\ &= \frac{1}{2}\int_\Omega b'(u)(\mathbf{x},t)w^2(\mathbf{x},t) + \frac{1}{2}\int_0^t \int_\Omega \big(b''(u)(\partial_t u)\big)(\mathbf{x},s)w^2(\mathbf{x},s)\,d\mathbf{x}ds,\end{aligned}$$

using the nonnegativity of b' and a Poincaré–Friedrichs inequality, we obtain

$$\begin{aligned}&\tfrac{\alpha}{2}\,\|w(\cdot,t)\|^2_{L^2(\Omega)} + \tfrac{1}{2}\int_0^t |w(\cdot,s)|^2_{H^1(\Omega)}\,ds \\ &\qquad \le \frac{1}{2}\int_0^t \int_\Omega \big(-b''(u)(\mathbf{x},s)\partial_t u(\mathbf{x},s) + (k\circ b)'^2(\mathbf{u})(\mathbf{x},s)\big)w^2(\mathbf{x},s)\,d\mathbf{x}ds \\ &\qquad\qquad + c\int_0^t \|F(\cdot,s)\|^2_{L^2(\Omega)},\end{aligned}$$

whence, thanks to Grönwall's lemma,

$$\|w(\cdot,t)\|^2_{L^2(\Omega)} \le c(u) \int_0^t \|F(\cdot,s)\|^2_{L^2(\Omega)}, \tag{5.8}$$

where all constants $c(u)$ only depend on u (we do not make them precise for simplicity). Next, taking v equal to $\partial_t w$ in (5.7), recalling that the function b' is nonnegative, we obtain

$$\begin{aligned}\alpha \int_0^t \|\partial_t w(\cdot,s)\|^2_{L^2(\Omega)}\,ds + |w(\cdot,t)|^2_{H^1(\Omega)} \\ \le \int_0^t \int_\Omega F(\mathbf{x},s)\partial_t w(\mathbf{x},s)\,d\mathbf{x}\,ds - \int_0^t \int_\Omega b''(u)(\mathbf{x},s)\partial_t u(\mathbf{x},s)\,w(\mathbf{x},s)\partial_t w(\mathbf{x},s)\,d\mathbf{x}\,ds \\ - \int_0^t \int_\Omega \big((k\circ b)'(u)w\big)(\mathbf{x},s)\,\cdot\,\nabla\partial_t w(\mathbf{x},s)\,d\mathbf{x}\,ds.\end{aligned}$$

Bounding the first term of the right-hand side is easy

$$\int_0^t \int_\Omega F(\mathbf{x},s)\partial_t w(\mathbf{x},s)\,d\mathbf{x}\,ds \le \frac{\alpha}{4}\int_0^t \|\partial_t w(\cdot,s)\|^2_{L^2(\Omega)}\,ds + \frac{1}{\alpha}\int_0^t \|F(\cdot,s)\|^2_{L^2(\Omega)}\,ds.$$

To handle the second term, we mainly use a Poincaré–Friedrichs inequality

$$\begin{aligned}\Big|\int_0^t \int_\Omega b''(u)(\mathbf{x},s)\partial_t u(\mathbf{x},s)\,w(\mathbf{x},s)\partial_t w(\mathbf{x},s)\,d\mathbf{x}\,ds\Big| \\ \le \frac{\alpha}{4}\int_0^t \|\partial_t w(\cdot,s)\|^2_{L^2(\Omega)}\,ds + c(u)\int_0^t |w(\cdot,s)|^2_{H^1(\Omega)}\,ds.\end{aligned}$$

Bounding the third term relies on an integration by parts with respect to t and also on the Poincaré–Friedrichs inequality:

$$\begin{aligned}\Big|\int_0^t \int_\Omega \big(k\circ b)'(u)w\big)(\mathbf{x},s)\,\cdot\,\nabla\partial_t w(\mathbf{x},s)\,d\mathbf{x}\,ds\Big| \le \frac{1}{2}|w(\cdot,t)|^2_{H^1(\Omega)} + \frac{c(u)}{2}\|w(\cdot,t)\|^2_{L^2(\Omega)} \\ + \frac{\alpha}{4}\int_0^t \|\partial_t w(\cdot,s)\|^2_{L^2(\Omega)}\,ds + c(u)\int_0^t |w(\cdot,s)|^2_{H^1(\Omega)}\,ds\end{aligned}$$

All this combined with (5.8) leads to

$$\begin{aligned}\frac{\alpha}{2}\int_0^t \|\partial_t w(\cdot,s)\|^2_{L^2(\Omega)}\,ds + \frac{1}{2}\,|w(\cdot,t)|^2_{H^1(\Omega)} \\ \le c(u)\int_0^t \|F(\cdot,s)\|^2_{L^2(\Omega)}\,ds + c'(u)\int_0^t |w(\cdot,s)|^2_{H^1(\Omega)}\,ds.\end{aligned}$$

By using the Grönwall's lemma, we conclude that

$$\alpha\int_0^t \|\partial_t w(\cdot,s)\|^2_{L^2(\Omega)}\,ds + |w(\cdot,t)|^2_{H^1(\Omega)} \le c(u)\int_0^t \|F(\cdot,s)\|^2_{L^2(\Omega)}\,ds. \tag{5.9}$$

2) Let $(\mathbb{H}_n)_n$ be an increasing sequence of finite-dimensional subspaces of $H^1_D(\Omega)$ such that $\cup_n \mathbb{H}_n$ is dense in $H^1_D(\Omega)$. It follows from the Cauchy–Lipschitz theorem [[24], Section 21] that problem (5.6) – (5.7) with $H_D(\Omega)$ replaced by $\mathbb{H}_n$ has a unique solution w_n in $\mathscr{C}^0(0,T;\mathbb{H}_n)\cap H^1(0,T;L^2(\Omega))$. Obviously, the sequence $(w_n)_n$ satisfies (5.9), so that there exists a subsequence, still denoted by $(w_n)_n$ for simplicity, which converges to a function w weakly in

$L^2(0,T;H^1_D(\Omega)) \cap H^1(0,T;L^2(\Omega))$. Thus, the function w is a solution of $(3.6)-(3.7)$ first for any v in a fixed $\mathbb{H}_m$, second for all v in $H^1_D(\Omega)$ by density. This yields the existence of a solution to problem $(5.6)-(5.7)$.
3) Let w be a solution of problem $(5.6)-(5.7)$ with data $F=0$. It follows from (5.8) that w is zero, which implies the uniqueness of the solution of problem $(5.6)-(5.7)$.

Even if the proof of this lemma is rather technical, it makes much simpler the remaining part of the estimates. For simplicity, we also introduce the fully discrete space, i.e, the space $\mathbb{W}_{N\tau}$ of functions which are affine on each interval $[t_{j-1},t_j]$, $1\le j\le J$, and such that their values in t_j belong to $\mathbb{X}_N$. It is readily checked that this space is finite-dimensional and imbedded in $\mathbb{W}$.

5.2 About the time discretization

From now on, we denote by u_τ the function which is affine on each interval $[t_{j-1},t_j]$ and equal to u^j at each time t_j, $0\le j\le J$, and also by $\mathbf{q}_\tau$ the function which is equal to $\mathbf{q}^j$ on each interval $]t_{j-1},t_j]$, $1\le j\le J$. For each function v continuous on $[0,T]$, we also introduce:
• the functions $\pi^+_\tau v$ and $\pi^-_\tau v$ which are constant, equal to $v(t_j)$ and $v(t_{j-1})$, respectively, on each interval $]t_{j-1},t_j]$, $1\le j\le J$,
• the function $i_\tau v$, equal to the piecewise affine Lagrange interpolate of v at the nodes t_j, $0\le j\le J$.

Let $\mathcal{T}_\tau$ denote the following semi-discrete operator: For any data F in $\mathscr{C}^0(0,T;L^2(\Omega))$ and (u_0,u_D) satisfying Assumption 3.1, $\mathcal{T}_\tau(F,u_0,u_D)$ is equal to the function $\widetilde{u_\tau}$ associated with the u^j solutions of

Find $(u^j)_{0\le j\le J}$ in $H^1(\Omega)^{J+1}$ such that

$$u^j = u_D(\cdot,t_j) \quad \text{on } \Gamma_D,\ 1\le j\le J, \qquad \text{and} \qquad u^0=u_0 \quad \text{in } \Omega, \tag{5.10}$$

and, for $1\le j\le J$,

$$\begin{aligned}\forall v\in H^1_D(\Omega),\quad \alpha\int_\Omega \Big(\frac{u^j-u^{j-1}}{\tau_j}\Big)(\mathbf{x})v(\mathbf{x})\,d\mathbf{x} + \int_\Omega \big(\nabla\cdot u^j\big)(\mathbf{x})\big(\nabla\cdot v\big)(\mathbf{x})\,d\mathbf{x}\\ = \int_\Omega F(\mathbf{x},t_j)v(\mathbf{x})\,d\mathbf{x}.\end{aligned} \tag{5.11}$$

We recall without proof three properties of this operator which are not completely standard, but easy to prove.
(i) Stability: For any data F in $\mathscr{C}^0(0,T;L^2(\Omega))$,

$$\|\mathcal{T}_\tau(F,0,0)\|_{\mathbb{W}} \le c\,\|\pi^+_\tau F\|_{L^2(0,T;L^2(\Omega))}. \tag{5.12}$$

(ii) A priori error estimate: For any data F in $\mathscr{C}^0(0,T.L^2(\Omega))$ such that $\mathcal{T}(F,u_0,u_D)$ belongs to $\mathscr{C}^1(0,T;H^1(\Omega))$,

$$\begin{aligned}&\|(\mathcal{T}-\mathcal{T}_\tau)(F,u_0,u_D,f)\|_{\mathbb{W}}\\ &\qquad\le c\,|\tau|^{\frac12}\,\big(\|\mathcal{T}(F,u_0,u_D)\|_{\mathscr{C}^1(0,T;H^1(\Omega))} + \|F\|_{\mathscr{C}^0(0,T;L^2(\Omega))}\big).\end{aligned} \tag{5.13}$$

(iii) Convergence: For any data F in $\mathscr{C}(0,T;L^2(\Omega))$,

$$\lim_{|\tau|\to0}\|(\mathcal{T}-\mathcal{T}_\tau)(F,0,0)\|_{\mathbb{W}} = 0. \tag{5.14}$$

Note that, since the Euler's scheme is of order 1, estimate (5.13) does not seem optimal, but it is because we are working in a stronger norm than the usual norm.

The second equation in (3.2) yields that the solution $(u^j, \mathbf{q}^j)_j$ of problem (3.1) − (3.2) satisfies $\mathbf{q}^j = -\nabla u^j - k \circ b(u^{j-1})\mathbf{e}_z$. Thus, problem (3.1) − (3.2) can equivalently be written

$$\mathcal{F}_\tau(u_\tau) = u_\tau - \mathcal{T}_\tau\big(\mathcal{G}_\tau(u), u_0, u_D\big) = 0, \tag{5.15}$$

where the mapping $\mathcal{G}_\tau$ is defined by

$$\begin{aligned}\int_\Omega \mathcal{G}_\tau(u)(\mathbf{x},t)v(x,t)\,d\mathbf{x}& \\ = -\langle \partial_t i_\tau\big(b(u)\big)(\cdot,t), v\rangle - \int_\Omega k \circ b(\pi_\tau^- u)(\mathbf{x},t)\mathbf{e}_z \,\cdot\, (\nabla v)(\mathbf{x},t)\,d\mathbf{x}.&\end{aligned} \tag{5.16}$$

5.3 About the space discretization

Similarly, we denote by $u_{N\tau}$ the function which is affine on each interval $[t_{j-1}, t_j]$ and equal to u_N^j at each time t_j, $0 \le j \le J$, and also by $\mathbf{q}_{N\tau}$ the function which is equal to $\mathbf{q}_N^j$ on each interval $]t_{j-1}, t_j]$, $1 \le j \le J$. It is readily checked that the function $u_{N\tau}$ belongs to the space $\mathbb{W}_{N\tau}$ introduced at the end of Section 5.1. We also define the discrete operator $\mathcal{T}_{N\tau}$: For any data F in $\mathscr{C}(0,T;L^2(\Omega))$ and (u_0, u_D) satisfying Assumption 4.1, $\mathcal{T}_{n\tau}(F, u_0, u_D)$ is equal to the function $\widetilde{u_{N\tau}}$ which interpolates the u_N^j solutions of

Find $(u_N^j)_{0\le j\le J}$ in $\mathbb{X}_N^{J+1}$ and $(\mathbf{q}_N^j)_{1\le j\le J}$ in $\mathbb{Y}_{NF}^J$ such that

$$u_N^0 = \mathcal{I}_{N-1}u_0 \quad \text{in } \Omega, \tag{5.17}$$

and, for $1 \le j \le J$,

$$\begin{aligned}&\forall w_N \in \mathbb{X}_N, \quad \alpha\Big(\frac{u_N^j - u_N^{j-1}}{\tau_j}, w_N\Big)_M + (\nabla\cdot\mathbf{q}_N^j, w_N)_M = \int_\Omega F(\mathbf{x},t_j)w_N(\mathbf{x})\,d\mathbf{x},\\ &\forall \varphi_N \in \mathbb{Y}_{NF}, \quad (\mathbf{q}_N^j, \varphi_N)_M - (u_N^j, \nabla\cdot\varphi_N)_M = -(u_{DN}^j, \varphi_N\cdot\mathbf{n})_M^{\Gamma_D}.\end{aligned} \tag{5.18}$$

There also, we state without proof some properties of this operator which can be derived from standard arguments in spectral methods, see [[6], Chap. V].
(i) Stability: For any data F in $\mathscr{C}(0,T;L^2(\Omega))$,

$$\|\mathcal{T}_{N\tau}(F,0,0)\|_{\mathbb{W}} \le c \sup_{v_N\in\mathbb{W}_{N\tau}} \frac{\displaystyle\sum_{j=1}^N \int_\Omega F(\mathbf{x},t_j)\big(v_N(\mathbf{x},t_j) - v_N(\mathbf{x},t_{j-1})\big)\,d\mathbf{x}}{\|v_N\|_{\mathbb{W}}}. \tag{5.19}$$

(ii) A priori error estimate: For any data F in $H^1(0,T.L^2(\Omega))$ such that $\mathcal{T}(F,u_0,u_D)$ belongs to $H^1(0,T;H^s(\Omega))$, $s > \frac{d}{2}$,

$$\|(\mathcal{T}_\tau - \mathcal{T}_{N\tau})(F,u_0,u_D)\|_{\mathbb{W}} \le c\,N^{1-s}\,\|\mathcal{T}(F,u_0,u_D)\|_{H^1(0,T;H^s(\Omega))}. \tag{5.20}$$

(iii) Convergence: For any data F in $\mathscr{C}(0,T;L^2(\Omega))$ and any τ,

$$\lim_{N\to+\infty} \|(\mathcal{T}_\tau - \mathcal{T}_{N\tau})(F,0,0)\|_{\mathbb{W}} = 0. \tag{5.21}$$

To conclude, we observe that problem (4.6) − (4.7) can equivalently be written as

$$\mathcal{F}_{N\tau}(u_{N\tau}) = u_{N\tau} - \mathcal{T}_{N\tau}\big(\mathcal{G}_{N\tau}(u_{N\tau}), u_0, u_D\big) = 0, \tag{5.22}$$

where the mapping $\mathcal{G}_{N\tau}$ is defined by

$$\begin{aligned}\int_\Omega \mathcal{G}_{N\tau}(u_{N\tau})(\mathbf{x},t)v_N(x,t)\,d\mathbf{x}\\ = -\langle \partial_t i_\tau\big(b(u_{N\tau})\big)(\cdot,t), v_N\rangle_M - (k\circ b(\pi_\tau^- u_{N\tau})(\mathbf{x},t)\mathbf{e}_z, \nabla v_N)_M\end{aligned} \tag{5.23}$$

(note that this new formulation requires once more the equivalence of the mixed discrete problem with a simpler one, which follows from the choice of the discrete spaces).

5.4 Some more lemmas

To go further, we introduce an approximation $u^*_{N\tau}$ of the solution u in $\mathbb{W}_{N\tau}$ and we make the following assumption.

Assumption 5.2.
(i) The solution u of problem (2.4) – (2.5) belongs to $L^\infty(\Omega\times]0,T[)$, together with its derivative $\partial_t u$;
(ii) The following convergence property holds

$$\lim_{|\tau|\to 0}\lim_{N\to+\infty} \|u - u^*_{N\tau}\|_{L^\infty(0,T;L^\infty(\Omega))} = 0. \tag{5.24}$$

Lemma 5.3. *Assume that the mappings b and k are of class $\mathscr{C}^2$ with bounded derivatives, and that the mapping b'' is Lipschitz-continuous. If Assumption 5.2 holds, there exist positive real numbers τ_0 and N_0 such that, for all τ, $|\tau|\le\tau_0$ and for all $N\ge N_0$, the operator $D\mathcal{F}_{N\tau}(u^*_{N\tau})$ is an isomorphism of $\mathbb{W}_{N\tau}$, and the norm of its inverse is bounded independently of τ and N.*

Proof: We use the expansion

$$D\mathcal{F}_{N\tau}(u^*_{N\tau}) = D\mathcal{F}(u) + \big(\mathcal{T}-\mathcal{T}_{N\tau}\big)\big(D\mathcal{G}(u),0,0\big) + \mathcal{T}_{N\tau}\big(D\mathcal{G}(u) - D\mathcal{G}_{N\tau}(u_{N\tau^*}),0,0\big).$$

From Lemma 5.1, $D\mathcal{F}(u)$ is an isomorphism of $\mathbb{W}$. So, we only have to check that the last two terms in the previous expansion tend to zero when $|\tau|$ and N^{-1} tend to zero.
1) We derive from (5.4) that, for any $w_{N\tau}$ in the unit sphere of $\mathbb{W}_{N\tau}$,

$$\begin{aligned}\int_\Omega D\mathcal{G}(u)(\mathbf{x},t)w_{N\tau}(\mathbf{x},t)v(x,t)\,d\mathbf{x} = -\langle b'(u)\partial_t w_{N\tau}(\cdot,t), v\rangle - \langle b''(u)w_{N\tau}(\mathbf{x},t)\partial_t u(\cdot,t), v\rangle\\ - \int_\Omega k\circ b(u)'(\mathbf{x},t)w_{N\tau}(\mathbf{x},t)\mathbf{e}_z\cdot\nabla v(\mathbf{x},t)\,d\mathbf{x}.\end{aligned}$$

Since $\mathbb{W}_{N\tau}$ is finite-dimensional, we deduce from Assumption 5.2 that $D\mathcal{G}(u)w_{N\tau}$ runs through a compact of $\mathcal{C}^0(0,T;L^2(\Omega))$. Thus, owing to the expansion

$$\mathcal{T}-\mathcal{T}_{N\tau} = \mathcal{T}-\mathcal{T}_\tau + \mathcal{T}_\tau - \mathcal{T}_{N\tau},$$

the convergence of the first term is a direct consequence of (5.14) and (5.21).
2) Owing to (5.19), we have to bound the three terms, for v_N in $\mathbb{W}_{N\tau}$ with $\|v_N\|_{\mathbb{W}}=1$,

$$\begin{aligned}E_1 &= -\langle b'(u)\partial_t w_{N\tau}(\cdot,t), v_N\rangle + (b'(u^*_{N\tau})\partial_t w_{N\tau}(\cdot,t), v_N)_M,\\ E_2 &= -\langle b''(u)w_{N\tau}(\mathbf{x},t)\partial_t u(\cdot,t), v_N\rangle + (b''(u^*_{N\tau})w_{N\tau}(\mathbf{x},t)\partial_t u(\cdot,t), v_N)_M\\ E_3 &= -\int_\Omega k\circ b(u)'(\mathbf{x},t)w_{N\tau}(\mathbf{x},t)\mathbf{e}_z\cdot\nabla v_N(\mathbf{x},t)\,d\mathbf{x}\\ &\qquad +(k\circ b(u^*_{N\tau}))'(\mathbf{x},t)w_{N\tau}(\mathbf{x},t)\mathbf{e}_z\cdot\nabla v_N)_M.\end{aligned}$$

To prove the convergence of E_1, we use the triangle inequality

$$E_1 = -\langle b'(u)\partial_t w_{N\tau}(\cdot,t), v_N\rangle + \langle b'(u^*_{N\tau})\partial_t w_{N\tau}(\cdot,t), v_N\rangle \\ +\langle b'(u^*_{N\tau})\partial_t w_{N\tau}(\cdot,t), v_N\rangle - (b'(u^*_{N\tau})\partial_t w_{N\tau}(\cdot,t), v_N)_M.$$

The convergence to 0 of the first line is a direct consequence of (5.24). To prove te convergence of the second one, we add and subtract any approximation $t_{M_\diamond}$ of $b'(u^*_{N\tau})$ and $v^*_{M_\diamond}$ of v_N in $\mathbb{P}_{M_\diamond}(\Omega)$, where $M_\diamond$ stands for the integer part of $\dfrac{2M-1-N}{2}$. Thus, triangle inequalities and the continuity property of the last term (see [[5], Eq. (13.28)] for instance) yields that the desired convergence follows from the convergence of the terms $\|b'(u^*_N) - \mathcal{I}_M b'(u^*_N)\|_{L^\infty(0,t;L^\infty(\Omega))}$, $\|b'(u^*_N) - t_{M_\diamond}\|_{L^\infty(0,t;L^\infty(\Omega))}$ and $\|v_N - v^*_{M_\diamond}\|_{L^2(\Omega)}$, so from the choice of $t_{M_\diamond}$ and $v^*_{M_\diamond}$. Bounding E_2 and E_3 relies on very similar arguments, that we omit for brevity.
This concludes the proof.

Lemma 5.4. *Assume that the mappings b and k are of class $\mathscr{C}^2$ with bounded derivatives, and that the mapping b'' is Lipschitz-continuous. If Assumption 5.2 holds, there exist a neighbourhood of $u^*_{N\tau}$ in $\mathbb{W}_N$ and a positive real number λ such that the following Lipschitz property holds for any z_N in this neigbourhood*

$$\|D\mathcal{F}_{N\tau}(u^*_{N\tau}) - D\mathcal{F}_{N\tau}(z_N)\|_{\mathcal{L}(\mathbb{W}_N)} \leq \lambda \|u_{N\tau^*} - z_N\|_{\mathbb{W}}, \tag{5.25}$$

where $\mathcal{L}(\mathbb{W}_N)$ stands for the space of endomorphisms of $\mathbb{W}_N$.

Proof: Thanks to (5.19), we are led to estimate the quantities, for all w_N and v_N in $\mathbb{W}_N$ with $\|v_N\|_{\mathbb{W}} = 1$,

$$L_1(t) = -(b'(u^*_{N\tau})\partial_t w_N(\cdot,t), v_N)_M + (b'(z_N)\partial_t w_N(\cdot,t), v_N)_M,$$

$$L_2(t) = -(b''(u^*_{N\tau})w_{N\tau}(\mathbf{x},t)\partial_t u^*_{N\tau}(\cdot,t), v_N)_M + (b''(z_N)w_{N\tau}(\mathbf{x},t)\partial_t z_N(\cdot,t), v_N)_M$$

$$L_3(t) = -\Big((k\circ b(u^*_{N\tau}))'(\cdot,t)w_{N\tau}(\cdot,t)\mathbf{e}_z, \nabla v_N\Big)_M + \Big((k\circ b(z_N))'(\cdot,t)w_{N\tau}(\cdot,t)\mathbf{e}_z, \nabla v_N\Big)_M.$$

To handle the quantity L_1, we observe that

$$|L_1(t)| \leq \|\partial_t w_N(\cdot,t)\|_{L^2(\Omega)} \|\mathcal{I}_M\big((b'(z_N) - b'(u^*_{N\tau})v_N\big)\|_{L^2(\Omega)}.$$

Thus, using once more [[5], Eq. (13.28)], the stability of $\mathcal{I}_M$, the Lipschitz-continuity of b' and the imbedding of $H^1(\Omega)$ into $L^4(\Omega)$, we deduce

$$\int_0^T |L_1(t)|\, dt \leq \|w_N\|_{\mathbb{W}} \|u^*_{N\tau} - z_N\|_{L^\infty(0,T;H^1(\Omega))}.$$

The term L_3 can be evaluated by exactly the same arguments, while bounding the term L_2 requires the addition and subtraction of a further term.

Lemma 5.5. *Assume that the mappings b and k are of class $\mathscr{C}^{\max\{\lceil s\rceil,2\}}$ with bounded derivatives. If the solution u belongs to $H^1(0,T;H^s(\Omega))$, $s > \frac{d+1}{2}$, the following estimate holds for the quantity $\varepsilon_{N\tau} = \|\mathcal{F}_{N\tau}(u^*_{N\tau})\|_{\mathbb{W}}$*

$$\varepsilon_{N\tau} \leq \|u - u^*_{N\tau}\|_{\mathbb{W}} + c(u)\,(|\tau|^{\frac{1}{2}} + N^{1-s}). \tag{5.26}$$

Proof: Since $\mathcal{F}(u) = 0$, we use the triangle inequality

$$\varepsilon_{N\tau} \le \|u - u^*_{N\tau}\|_{\mathbb{W}} + \|(\mathcal{T} - \mathcal{T}_\tau)(\mathcal{G}(u), u_0, u_D)\|_{\mathbb{W}} + \|(\mathcal{T}_\tau - \mathcal{T}_{N\tau})(\mathcal{G}(u), u_0, u_D)\|_{\mathbb{W}} + \|\mathcal{T}_{N\tau}(\mathcal{G}_(u) - \mathcal{G}_{N\tau}(u_{N\tau^*}, 0, 0)\|_{\mathbb{W}}.$$

Evaluating the first three terms follows from (5.13) and (5.20) (indeed, the fact that $\mathcal{G}(u)$ belongs to $\mathscr{C}^0(0,T;L^2(\Omega))$ is easily derived from the regularity properties of u). To handle the last term, owing to (5.19), we only have to bound the quantities, for v_N running through the unit sphere of $\mathbb{W}_{N\tau}$,

$$G_1 = \sum_{j=1}^{J}\Big(-\langle \partial_t b(u)(\cdot,t_j), v_N(\cdot,t_j) - v_N(\cdot,t_{j-1})\rangle + \big(\partial_t i_\tau\big(b(u^*_{N\tau})\big)(\cdot,t), v_N(\cdot,t_j) - v_N(\cdot,t_{j-1})\big)_M\Big),$$

$$G_2 = \sum_{j=1}^{J}\Big(-\int_\Omega k\circ b(u)(\mathbf{x},t)\mathbf{e}_z \cdot \nabla\big(v_N(\cdot,t_j) - v_N(\cdot,t_{j-1})\big)\,d\mathbf{x} + \big(k\circ b(\pi^-_\tau u_{N\tau})(\mathbf{x},t)\mathbf{e}_z, \nabla\big(v_N(\cdot,t_j) - v_N(\cdot,t_{j-1})\big)\big)_M\Big).$$

By combining the regularity properties of u with [[7], Thm 1], we obtain the desired estimate.

Due to the regularity assumed on u in the previous lemma , the function $u^*_{N\tau}$ can easily be chosen such that

$$\|u - u^*_{N\tau}\|_{\mathbb{W}} \le c(u)\,(|\tau|^{\frac{1}{2}} + N^{1-s}), \tag{5.27}$$

which yields the convergence of $\mathcal{F}_{N\tau}(u^*_{N\tau})$ to zero.

5.5 The conclusive a priori error estimates

Owing to Lemmas 5.3 to 5.5, all the assumptions needed to apply the theorem of Brezzi, Rappaz and Raviart [[8], Thm 1] (see also [[14], Chap. IV, Thm 3.1]) are satisfied. This yields the estimate for $\|u - u_{N\tau}\|_{\mathbb{W}}$. The estimate for $\|\mathbf{q} - \mathbf{q}_{N\tau}\|_{L^2(0,T;H(\mathrm{div},\Omega))}$ is then easily derived by hand. So, we now state the main result of this section.

Theorem 5.6. *Assume that, for some real number $s > \frac{d+1}{2}$,*
(i) the mappings b and k are of class $\mathscr{C}^{\max\{\lceil s\rceil,2\}}$ with bounded derivatives,
(ii) the solution u of problem (2.4) – (2.5) belongs to $H^2(0,T;L^2(\Omega)\cap H^1(0,T;H^s(\Omega))$.
Thus, there exist positive real numbers τ_0^* and N_0^* such that, for all τ, $|\tau| \le \tau_0^*$ and for all $N \ge N_0^*$, the solution $(u^j_N, \mathbf{q}^j_N)$ of problem (4.6) – (4.7) satisfies the following a priori error estimates:

$$\|u - u_{N\tau}\|_{\mathbb{W}} + \|\mathbf{q} - \mathbf{q}_{N\tau}\|_{L^2(0,T;H(\mathrm{div},\Omega))} \le c(u)\,(|\tau|^{\frac{1}{2}} + N^{1-s}), \tag{5.28}$$

where the constant $c(u)$ only depends on the solution u of problem (2.4) – (2.5).

Despite the technicity of its proof, this estimate seems fully optimal. Note that the use of a discretization relying on the mixed formulation is only justified by the fact that $\mathbf{q}$ is a physical variable.

6 An iterative algorithm

We propose here an iterative scheme to solve the nonlinear problems resulting from the discretization procedure, following the approach in [17], and we prove the convergence of this scheme.

To define the scheme, we note from part (i) of Assumption 2.1, that there exists a positive real K_b such that $0 \leq b'(\xi) \leq K_b$, for all real numbers ξ. We set: $b_\alpha(\xi) = \alpha\xi + b(\xi)$. Let $K \geq K_b$ be a real constant. We fix an integer $j \in \{1, \cdots, J\}$ and start with $u_N^{j,0} = u_N^{j-1}$ (or an approximation of it). For a given $i > 0$. assuming that u_N^{j-1}, $u_N^{j,i-1}$ are known in $\mathbb{X}_N$ and that $\mathbf{q}_N^{j-1}$ and $\mathbf{q}_N^{j,i-1}$are known in $\mathbb{Y}_N$, we consider the following linear problem

Find $u_N^{j,i}$ in $\mathbb{X}_N$ and $\mathbf{q}_N^{j,i}$ in $\mathbb{Y}_N$ such that

$$\mathbf{q}_N^{j,i} \cdot \mathbf{n} = -i_{N-1}^{\Gamma_F} f(\cdot, t_j) \quad \text{on } \Gamma_F \tag{6.1}$$

and

$$\begin{aligned} \forall w_N \in \mathbb{X}_N, \quad & K(u_N^{j,i}, w_N)_M + \tau_j(\nabla \cdot \mathbf{q}_N^{j,i}, w_N)_M \\ & = (K u_N^{j,i-1} + b_\alpha(u_N^{j-1}) - b_\alpha(u_N^{j,i-1}), w_N)_M, \\ \forall \varphi_N \in \mathbb{Y}_{NF}, \quad & (\mathbf{q}_N^{j,i}, \varphi_N)_M - (u_N^{j,i}, \nabla \cdot \varphi_N)_M \\ & = -\big(\mathcal{I}_{N-1}\big(k \circ b(u_N^{j-1}) e_z\big), \varphi_N\big)_M - (u_{DN}^j, \varphi_N \cdot \mathbf{n})_M^{\Gamma_D}. \end{aligned} \tag{6.2}$$

The same arguments as for problem (4.6)−(4.7) lead to prove the well-posedness of this problem. Moreover, solving it is not expensive. We stop the iteration on i arbitrarily, for instance when $u_N^{j,i} - u_N^{j,i-1}$ becomes smaller than a given tolerance.

To study the convergence of this scheme, we use the following notation

$$e_u^i = u_N^j - u_N^{j,i}, \qquad \mathbf{e}_\mathbf{q}^i = \mathbf{q}_N^j - \mathbf{q}_N^{j,i}, \tag{6.3}$$

where u_N^j and $\mathbf{q}_N^j$, $1 \leq j \leq J$, are the solution of problem (4.6) − (4.7). We first recall from [[2], Lemma 3.1] the next inf-sup condition result.

Lemma 6.1. *For any w_N in $\mathbb{X}_N$, there exists a φ_N in $\mathbb{Y}_{NF}$ such that*

$$\nabla \cdot \varphi_N = w_N \qquad \textit{and} \qquad \|\varphi_N\|_{L^2(\Omega)^d} + \|\nabla \cdot \varphi_N\|_{L^2(\Omega)} \leq c_\diamond \|w_N\|_{L^2(\Omega)}. \tag{6.4}$$

We are now in a position to prove the following convergence result:

Theorem 6.2. *There exists a positive constant $c_\sharp < 1$ such that, for each $i > 0$,*

$$\|e_u^i\|_{L^2(\Omega)} \leq c_\sharp^i \|e_u^0\|_{L^2(\Omega)} \qquad \textit{and} \qquad \|\mathbf{e}_\mathbf{q}^i\|_{L^2(\Omega)^d} \leq c_\sharp^{i-1} \, \frac{K + \tau_j/9c_\diamond}{2K\tau_j^{1/2}} \, \|e_u^0\|_{L^2(\Omega)}, \tag{6.5}$$

for all j, $1 \leq j \leq J$.

Proof: We prove successively the two estimates in (6.5).
1) By noting that the first equation in (4.7) can be written

$$(b_\alpha(u_N^j) - b_\alpha(u_N^{j-1}), w_N)_M + \tau_j(\nabla \cdot q_N^j, w_N)_M = 0,$$

subtracting from it the first equation in (6.2) and noting that $u_N^{j,i} - u_N^{j,i-1} = -\big(e_u^i - e_u^{i-1}\big)$, we derive

$$K(e_u^i - e_u^{i-1}, w_N)_M + \tau_j(\nabla \cdot \mathbf{e}_\mathbf{q}^i, w_N)_M + (b_\alpha(u_N^j) - b_\alpha(u_N^{j,i-1}), w_N)_M = 0. \tag{6.6}$$

Taking $w_N = e_u^i$ yields

$$K\|e_u^i\|_M^2 + \tau_j(\nabla \cdot \mathbf{e}_\mathbf{q}^i, e_u^i)_M = (K e_u^{i-1} - b_\alpha(u_N^j) + b_\alpha(u_N^{j,i-1}), e_u^i)_M, \tag{6.7}$$

where we denote by $\|w\|_M$ the quantity $(w,w)_M^{1/2}$, for all continuous function w. By subtracting the second equation in (6.2) from the second equation in (4.7), and since the convection term is discretized explicitly, we have

$$(\mathbf{e}_{\mathbf{q}}^i, \varphi_N)_M - (e_u^i, \nabla \cdot \varphi_N)_M = 0. \tag{6.8}$$

Note that the discrete products in (6.7) can be replaced by the scalar product in $L^2(\Omega)$, due to the exactness property (4.4). By Lemma 6.1, there exists a φ_N in Y_{NF} such that $\nabla \cdot \varphi_N = e_u^i$ and

$$\|\varphi_N\|_{L^2(\Omega)} \leq c_\diamond \|e_u^i\|_{L^2(\Omega)}.$$

Using this φ_N in (6.8), together with the inequality of Cauchy–Schwarz yields

$$\|e_u^i\|_{L^2(\Omega)} \leq c_\diamond \|\mathbf{e}_{\mathbf{q}}^i\|_{L^2(\Omega)^d}. \tag{6.9}$$

Now, taking $\varphi_N = \tau_j \mathbf{e}_{\mathbf{q}}^i$ in (6.8) and adding the result to (6.7) gives

$$K\|e_u^i\|_M^2 + \tau_j \|\mathbf{e}_{\mathbf{q}}^i\|_M^2 = (Ke_u^{i-1} - b_\alpha(u_N^j) + b_\alpha(u_N^{j,i-1}), e_u^i)_M \tag{6.10}$$

Since $\alpha \leq b'_\alpha(\xi) \leq K_b \leq K$, for all real numbers ξ, it follows that

$$|Ke_u^{i-1} - (b_\alpha(u_N^j) - b_\alpha(u_N^{j,i-1}))| \leq |(K-\alpha)e_u^{i-1}| \tag{6.11}$$

Using (4.4), the Cauchy–Schwarz inequality, (6.9) and the last inequality leads to

$$(K + \frac{\tau_j}{9c_\diamond})\|e_u^i\|_{L^2(\Omega)}^2 \leq (K-\alpha)\|e_u^{i-1}\|_{L^2(\Omega)}\|e_u^i\|_{L^2(\Omega)}.$$

All this yields

$$\|e_u^i\|_{L^2(\Omega)} \leq c_\sharp \|e_u^{i-1}\|_{L^2(\Omega)},$$

where the constant $c_\sharp = \frac{(K-\alpha)}{K+\tau_j/9c_\diamond}$ is < 1. The first part of (6.5) is, then, proven.
2) For the second inequality, from (6.10) and (6.11) and again (4.4) we obtain

$$K\|e_u^i\|_{L^2(\Omega)}^2 + \tau_j \|\mathbf{e}_{\mathbf{q}}^i\|_{L^2(\Omega)^d}^2 \leq (K-\alpha)\|e_u^{i-1}\|_{L^2(\Omega)}\|e_u^i\|_{L^2(\Omega)}$$

Using, now, the inequality $|ab| \leq \delta a^2 + b^2/4\delta^2$, for all reals a and b and $\delta > 0$, yields

$$(K-\alpha)\|e_u^{i-1}\|_{L^2(\Omega)}\|e_u^i\|_{L^2(\Omega)} \leq K\|e_u^i\|_{L^2(\Omega)}^2 + \frac{(K-\alpha)^2}{4K^2}\|e_u^{i-1}\|_{L^2(\Omega)}.$$

So, we derive

$$\|\mathbf{e}_{\mathbf{q}}^i\|_{L^2(\Omega)} \leq \frac{(K-\alpha)}{2K\tau_j^{1/2}}\|e_u^{i-1}\|_{L^2(\Omega)}.$$

The estimate for e_u^i then gives the second part of (6.5).

Theorem 6.2 shows that, in the case where the convection is discretized explicitly, the iterative scheme (6.1)-(6.2) is convergent in both pressure and flux, in particular, in absence of convection ($k = 0$). Moreover it follows from (6.5) that the convergence is geometric hence very fast.

Remark 6.3. If the discrete problem (4.6) – (4.7) was constructed without explicitation of the last term, i.e., with the term $(\mathcal{I}_{N-1}(k \circ b(u_N^{j-1})\mathbf{e}_z)$ replaced by $(\mathcal{I}_{N-1}(k \circ b(u_N^{j})\mathbf{e}_z)$, a linearization procedure could also be applied to solve the discrete problem. The corresponding linear problem there reads

Find $u_N^{j,i}$ in $\mathbb{X}_N$ and $\mathbf{q}_N^{j,i}$ in $\mathbb{Y}_N$ such that

$$\mathbf{q}_N^{j,i} \cdot \mathbf{n} = -i_{N-1}^{\Gamma_F} f(\cdot, t_j) \quad \text{on } \Gamma_F \tag{6.12}$$

and

$$\begin{array}{l} \forall w_N \in \mathbb{X}_N, \\ K(u_N^{j,i}, w_N)_M + \tau_j(\nabla \cdot \mathbf{q}_N^{j,i}, w_N)_M = (K u_N^{j,i-1} + b_\alpha(u_N^{j-1}) - b_\alpha(u_N^{j,i-1}), w_N)_M, \\ \forall \varphi_N \in \mathbb{Y}_{NF}, \quad (\mathbf{q}_N^{j,i}, \varphi_N)_M - (u_N^{j,i}, \nabla \cdot \varphi_N)_M \\ \qquad\qquad = -\big(\mathcal{I}_{N-1}\big(k \circ b(u_N^{j,i-1})e_z\big), \varphi_N\big)_M - (u_{DN}^{j}, \varphi_N \cdot n)_M^{\Gamma_D}. \end{array} \tag{6.13}$$

However, this problem is a little more complex than (6.1) – (6.2). Moreover, the convergence of the iterative algorithm is only proved with a weak restriction, namely when $|\tau|$ is small enough. This confirms that our choice of discretization is more convenient for the final implementation.

7 Some numerical experiments

In this section, we present some numerical results in dimension $d = 2$ for the simulation of Richards equation in its form after the Kirchhoff transformation (2.1).These experiments are compared to the theoretical convergence results provided in Section 6. These tests are made first on a convex domain. The extension to more general domains is presented in Section 8.

Concerning the time scheme, we use the backward Euler discretization with uniform time step $\tau_n = \delta t$. The obtained linear systems are solved using a preconditioned GMRES (Generalized Minimal Residual) iterative routine, see for instance Saad [23]. Note that, all the computations have been performed using FreeFEM3D–spectral version developed during the Ph.D Thesis of Yakoubi, see [30] and also [9].

We consider the model domain $\Omega =]-1, 1[^2$ and the final time $T = 1$, with Γ_F is equal to the top of the domain $\{y = 1\}$, $\Gamma_D = \partial\Omega \setminus \Gamma_F$, the coefficients: $\alpha = 1$, $k = 1$ and the function $b(s) = \dfrac{s^3}{s^2+1}$. At $t = 0$, the initial condition $u_0(x, y)$ is equal to 0, the Dirichlet condition on u on Γ_D is given by : $u_D(x, -1) = u_D(-1, y) = u_D(1, y) = 0$ and we consider suitable forcing function f corresponding to the exact solution :

$$u(x, y; t) = \cos(\pi x)\, \sin(\pi y)\, t. \tag{7.1}$$

Note that, by using this solution, the error du to the time discretization (Euler scheme) is neglected compared with the space error. We fix the time step δt equal to 0.1 and we plot the errors between the numerical solution $(u, \mathbf{q})$ and the exact one, in norms $L^2(\Omega)$, $H^1(\Omega)$ and $H(\mathrm{div}, \Omega)$, at the final time $T = 1$ by varying the polynomial degree from $N = 5$ to $N = 25$. Note that $\mathbf{q}$ is computed thanks to formula (2.11) and (7.1), see Figures 1. and 2.

All these results are in good coherence with the estimates proved in Section 6. They confirm the efficiency of the spectral method for solving the nonlinear problem (2.1).

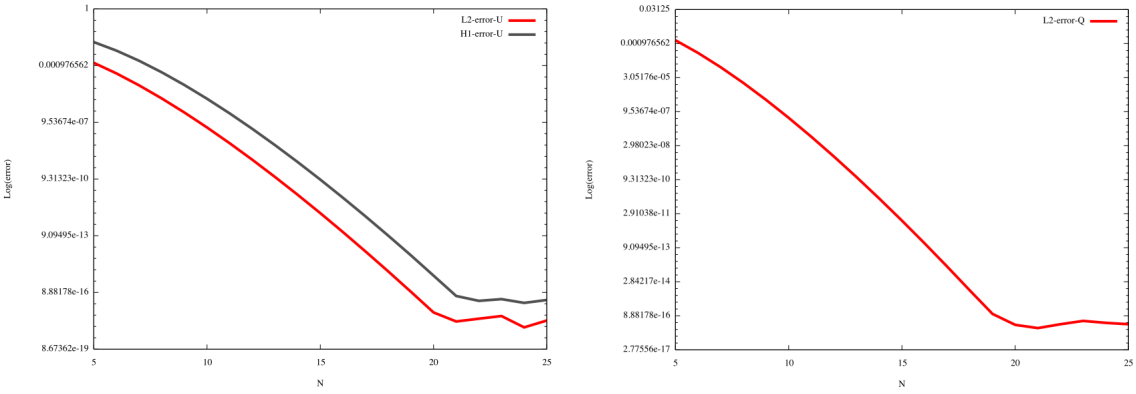

Figure 1: Error curves as a function of N, on u in norm $L^2(\Omega)$ and $H^1(\Omega)$ (left), and on flux $\mathbf{q}$ in norm $L^2(\Omega)$ (right).

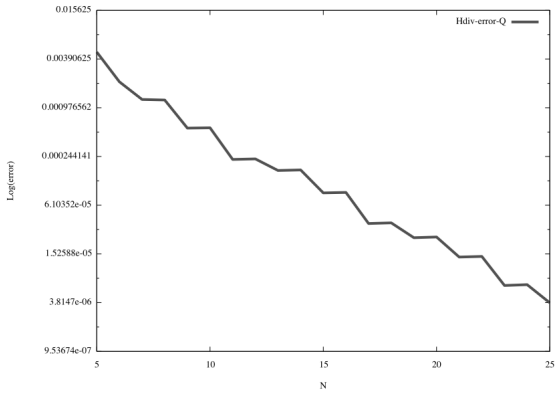

Figure 2: Error curve on $\mathbf{q}$ in norm $H(\mathrm{div}, \Omega)$.

We represent now the isovalues curves of u, $\mathbf{q}_x$ and $\mathbf{q}_y$, the x-component and the y-component of the flux $\mathbf{q}$, respectively, at the final time $T = 1$ and with a polynomial degree N equal to 15, see Figures 3, 4 and 5 for the exact solutions and the computed ones.

In Table 1, we present the errors at the final time $T = 1$, namely

$$E_u = \|u - u_{N\tau}\|_{L^2(\Omega)}, \quad E_{1,u} = \|u - u_{N\tau}\|_{H^1(\Omega)}, \quad E_{\partial_t u} = \|\partial_t u - \partial_t u_{N\tau}\|_{L^2(\Omega)}$$

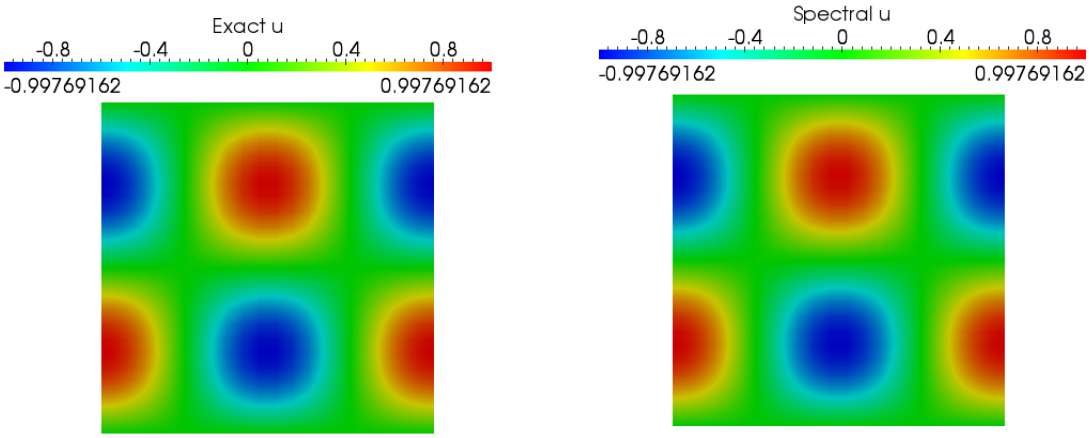

Figure 3: Isovalues of the the exact solution (7.1) at the final time (left) and of the spectral

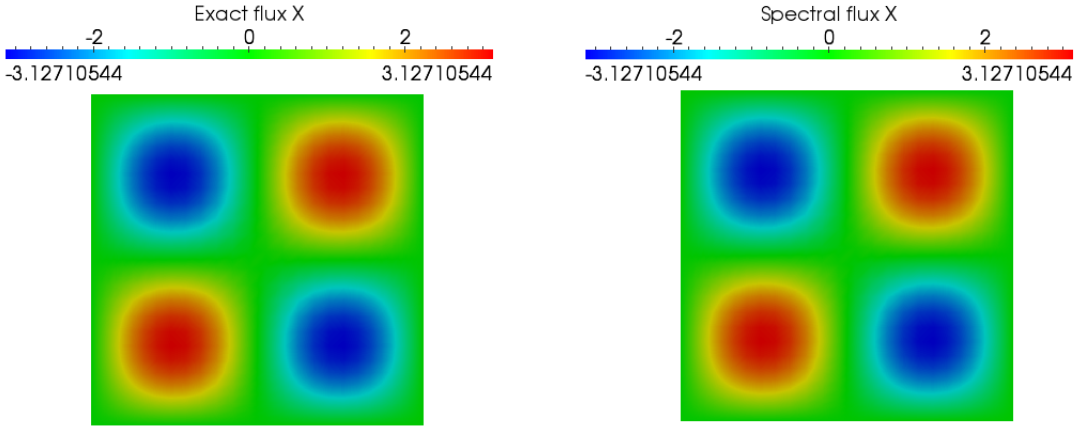

Figure 4: Isovalues of the component $\mathbf{q}_x$ of $\mathbf{q}$, the exact solution (left) and the spectral one (right).

and

$$E_q = \|\mathbf{q} - \mathbf{q}_{N\tau}\|_{L^2(\Omega)}, \quad E_{div,q} = \|\mathbf{q} - \mathbf{q}_{N\tau}\|_{H(div,\Omega)}.$$

Table 2 presents the errors E_u, $E_{1,u}$ and $E_{\partial_t u}$ for decreasing time steps.

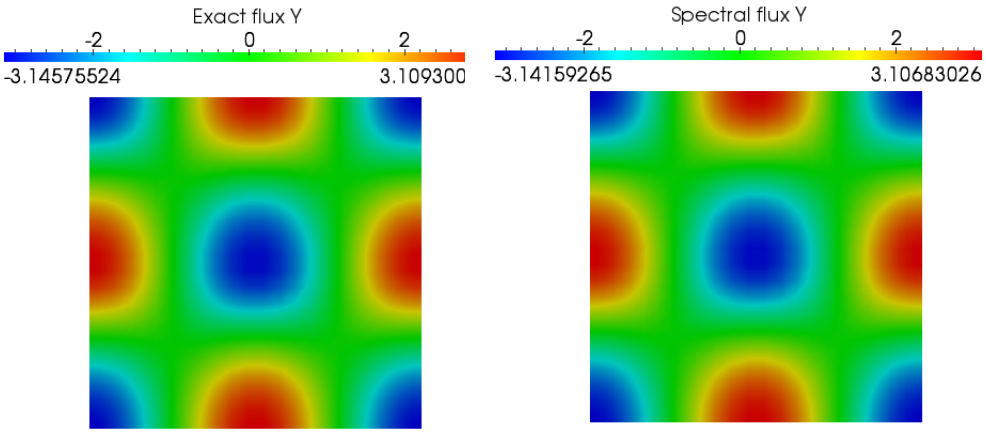

Figure 5: Isovalues of the component $\mathbf{q}_y$ of $\mathbf{q}$, the exact solution (left) and the spectral one (right).

N	E_u	$E_{1,u}$	$E_{\partial_t u}$	E_{q}	$E_{div,\mathrm{q}}$
5	5.59605e-05	0.0014327	5.59605e-05	0.00030507	0.00690947
6	7.42798e-06	0.000244397	7.42798e-06	0.000249393	0.00679991
7	8.42475e-07	3.43835e-05	8.42475e-07	7.99485e-05	0.00311564
8	8.34261e-08	4.11144e-06	8.34261e-08	7.99251e-05	0.00311508
9	7.33169e-09	4.27247e-07	7.33169e-09	2.44149e-05	0.00135193
10	5.79206e-10	3.9245e-08	5.79206e-10	2.44223e-05	0.00135179
11	4.15598e-11	3.22939e-09	4.15598e-11	8.12011e-06	0.00056492
12	2.73158e-12	2.40645e-10	2.73158e-12	8.12263e-06	0.00056488
13	1.65633e-13	1.63836e-11	1.65634e-13	2.65983e-06	0.000229697
14	9.32148e-15	1.02666e-12	9.33269e-15	2.6607e-06	0.000229687
15	4.89239e-16	5.9576e-14	5.82667e-16	8.86705e-07	9.14802e-05
16	2.67924e-17	3.26966e-15	2.66312e-16	8.86999e-07	9.14783e-05
17	5.16786e-17	3.68451e-16	5.29578e-16	2.95653e-07	3.58423e-05
18	1.13366e-17	3.83362e-16	2.16114e-16	2.95753e-07	3.58424e-05
19	5.90463e-17	3.95774e-16	1.15383e-15	9.89926e-08	1.38593e-05
20	2.39087e-17	4.7715e-16	2.88674e-16	9.90265e-08	1.38596e-05

Table 1: The errors for increasing polynomial degree N.

Table 2 shows that the convergence rate is 1, whereas the theoretical result above shows that the rate order is equal to $\frac{1}{2}$. This is a simple consequence of using the usual norms in our computations while in estimate (5.28), we use the strong norms.

τ	E_u	$E_{1,u}$	$E_{\partial_t u}$
0.1	0.015255	0.123552	0.425451
0.05	0.00411098	0.0313272	0.219896
0.0334	0.00197304	0.0158733	0.145553
0.025	0.00127113	0.0111614	0.107691
0.02	0.000949681	0.00885111	0.0848985
0.0167	0.000765871	0.00737193	0.0697475
0.0143	0.000645066	0.00630424	0.0589909
0.0125	0.000558555	0.00548795	0.0509869
0.0111	0.000493454	0.00483903	0.0448424
0.01	0.000441041	0.00432276	0.0398966
0.00625	0.000273011	0.00256908	0.0238908
0.003125	0.000135058	0.00126857	0.0117644

Table 2: The errors for decreasing time steps.

8 Extension to the Spectral Element Approximation

Our attention here is only focussed to treat complex geometries, where the spectral method based on a single domain is no longer suitable. To handle these complex geometries, we use the spectral element method which combines the domain partition with the high accuracy of the spectral method.

Assuming now that Ω is the union without overlap of a finite number of rectangles ($d = 2$) or rectangular parallelepipeds ($d = 3$) Ω_k, $1 \leq k \leq K$, and that the intersection of two different Ω_k is either a vertex or an edge or a face of both of them, we define the discrete spaces for $k \in \{1, ..., K\}$

$$\mathbb{X}_N = \{v_N \in L^2(\Omega);\ v_N|_{\Omega_k} \in \mathbb{P}_{N-1}(\Omega_k)\}$$

$$\mathbb{Y}_N = \{q_N \in H(div, \Omega);\ q_N|_{\Omega_k} \in \mathbb{P}_{N,N-1}(\Omega_k) \times \mathbb{P}_{N-1,N}(\Omega_k)\} \text{ in } d = 2$$

and where $d = 3$

$$\mathbb{Y}_N = \{q_N \in H(div, \Omega);\ q_N|_{\Omega_k} \in \mathbb{P}_{N,N-1,N-1}(\Omega_k) \times \mathbb{P}_{N-1,N,N-1}(\Omega_k) \times \mathbb{P}_{N-1,N-1,N}(\Omega_k)$$

Denoting by F_k one of the affine mappings that maps $]-1, 1[^d$ on Ω_k, we now set the discrete product

$$(\varphi,\psi)_M^k = \begin{cases} \dfrac{meas(\Omega_k)}{4}\displaystyle\sum_{i,j=0}^{M} \varphi\circ F_k(\xi_{Mi},\xi_{Mj})\psi\circ F_k(\xi_{Mi},\xi_{Mj})\rho_{Mi}\rho_{Mj} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad d=2 \\ \dfrac{meas(\Omega_k)}{8}\displaystyle\sum_{i,j,k=0}^{M} \varphi\circ F_k(\xi_{Mi},\xi_{Mj},\xi_{Mk})\psi\circ F_k(\xi_{Mi},\xi_{Mj},\xi_{Mk})\rho_{Mi}\rho_{Mj}\rho_{Mk} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if} \quad d=3 \end{cases}$$

and finally,

$$(\varphi,\psi) = \sum_{k=1}^{K}(\phi,\psi)_M^k$$

Assuming that Γ_D is the union of whole edges ($d=2$) or faces ($d=3$) of the Ω_k, we can also define a discrete product $(.,.)_M^{\Gamma_D}$ on Γ_D in an obvious way and a Lagrange interpolation operator $i_M^{\Gamma_F}$ on Γ_F. With this new notation, the discrete problem is exactly the same as (4.6) − (4.7).

This yields that the discretization by the spectral element method is fully conforming. Exactly the same arguments as previously (see Section 6) lead to the analogue of estimate (5.28) in this case. Moreover, the regularity which is required for the solution is now local:

$$\|u-u_{N\tau}\|_{\mathbb{W}} + \|\mathbf{q}-\mathbf{q}_{N\tau}\|_{L^2(0,T;H(\mathrm{div},\Omega))} \leq \sum_{k=1}^{K} c_k(u)\,(|\tau|^{\frac{1}{2}} + N^{1-s_k}). \tag{8.1}$$

8.1 Space accuracy

The first test used to validate the space accuracy. We consider Richard's equation in two-dimensional domain Ω described in Figure 6 and suitable forcing functions such that the exact solution is given by (7.1).

The Dirichlet condition on u is prescribed on the boundary $\Gamma_D = \partial\Omega/\Gamma_F$, where Γ_F equal to the top of domain $y = \frac{1}{2}$. At $t = 0$, the initial condition $u_0(x, y)$ is equal to 0. We fix the time step δt equal to 0.1, and we plot the error in some norms between the numerical solution and the exact solution at the final time $t = 1$, with a successive polynomial degree from $N = 5$ to $N = 20$.

It is clear from Figures 7, 8 and 9 that the convergence errors coincide with the slope of the function e^{-x}. Hence the spectral convergence is obtained which is consistent with the error estimate (5.28).

In figures 10 and 11, we present the exact and spectral solutions at $T = 1$, when polynomial degree N equal to 15 in each sub-domain Ω_k. The results are similar and in total concordance with Figures 7, 8 and 9.

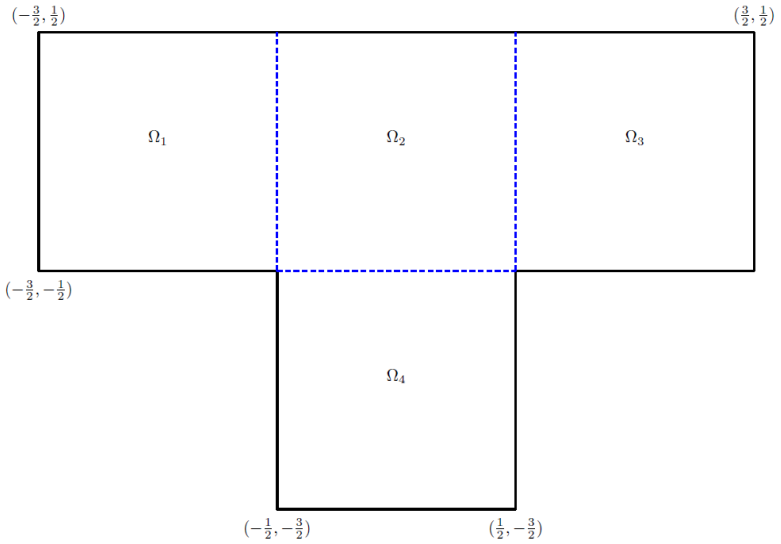

Figure 6: The computational domain.

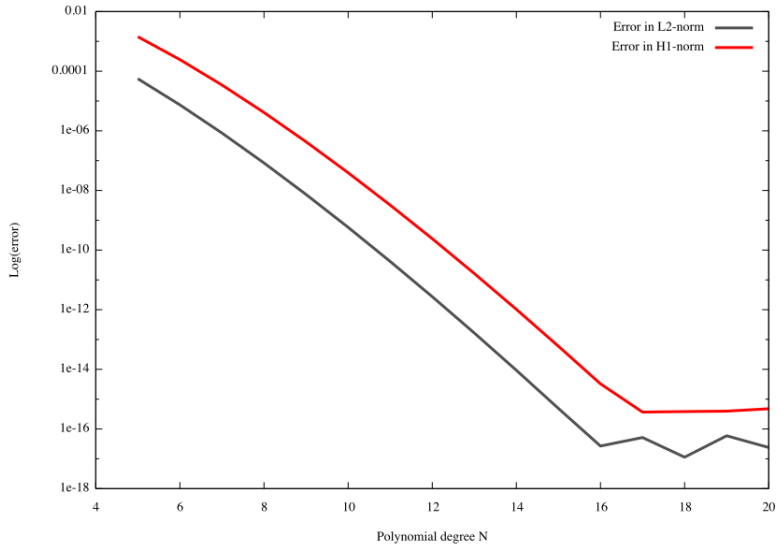

Figure 7: The L^2-errors and H^1-error on u as a function of N.

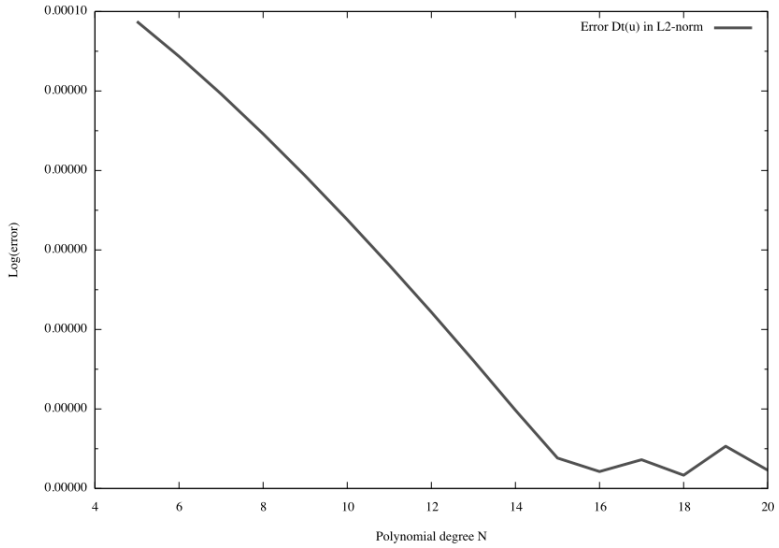

Figure 8: The L^2-errors on $\partial_t u$ as a function of N.

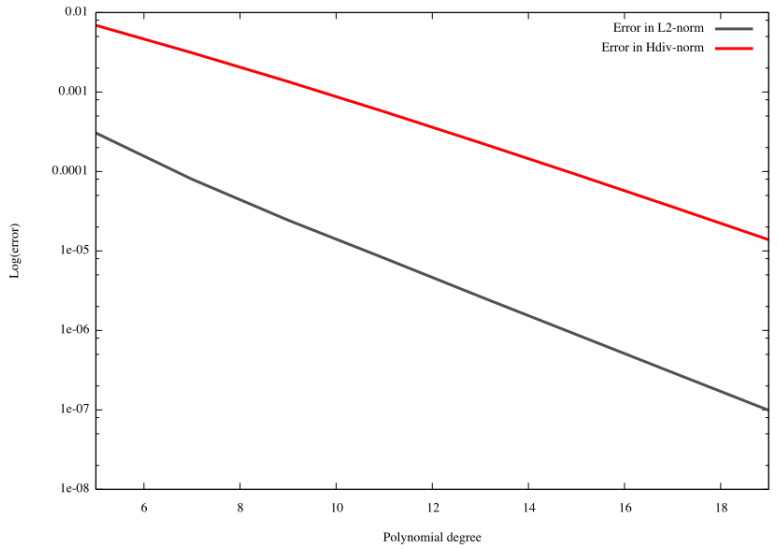

Figure 9: The L^2-errors and H_{div}-errors on flux $\mathbf{q}$ as a function of N.

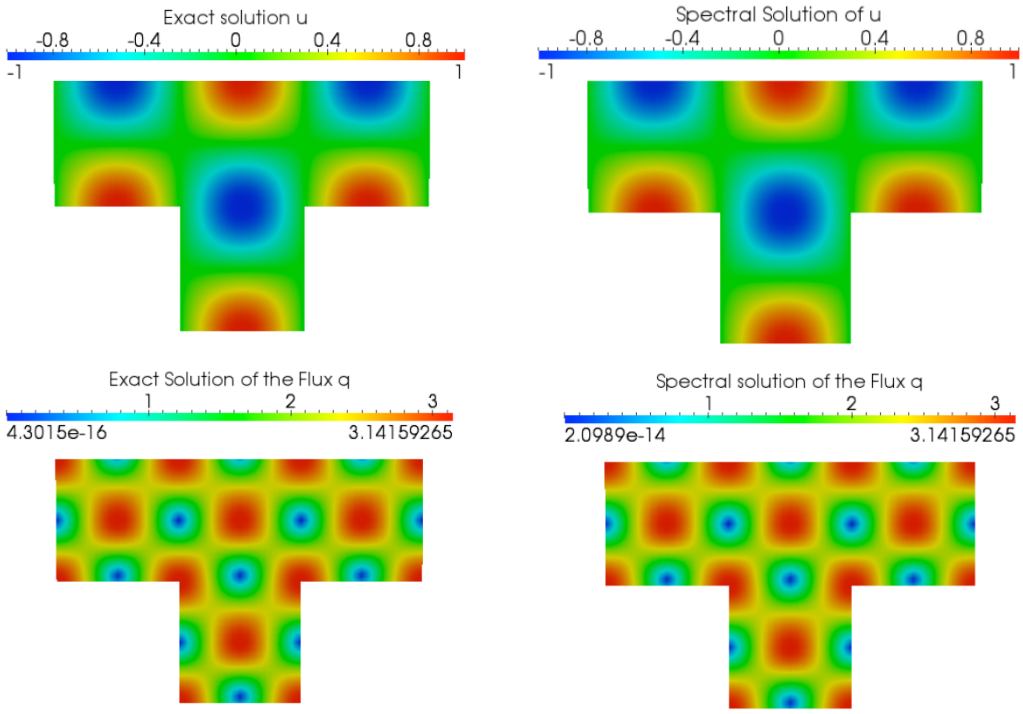

Figure 10: Exact solution (left) versus spectral element solution for $N = 15$ (right).

8.2 Time accuracy

The aim of the second test is to verify the time accuracy. We solve Richard's equation (2.1) in three-dimensional domain (see Figure 11.)

$$\Omega = \bigcup_{k=1}^{4} \Omega_k, \ where \qquad \begin{cases} \Omega_1 =]0, 0.5[\times]-0.5, 0[\times]0, 0.5[, \\ \Omega_2 =]0.5, 1[\times]-0.5, 0[\times]0, 0.5[, \\ \Omega_3 =]1, 1.5[\times]-0.5, 0[\times]0, 0.5[, \\ \Omega_4 =]0.5, 1[\times]-0.5, 0[\times]-0.5, 0[. \end{cases}$$

The exact solution is given by:

$$u_e(x, y, z; t) = \Big(\log(t + 1) + \cos(\pi\, t)\Big)\,(x^2 + y^2 + z^2).$$

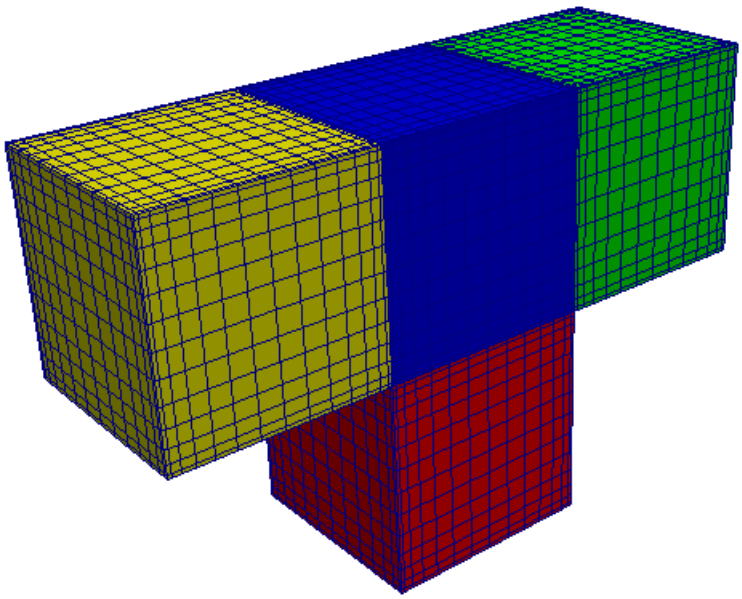

Figure 11: Computational domain with spectral grids.

In this case, the boundary Γ_F coincide with the top of the domain $\Gamma_F = \{z = 0\}$. On $\Gamma_D = \partial\Omega/\Gamma_F$, the Dirichlet condition is considered by taking $u(x, y, z; t) = u_e(x, y, z; t)$, and the initial value $u_0(x, y, z) = x^2 + y^2 + z^2$.

To verify the theoretical estimates, we have started performing computations using polynomial degree equal to 3 in each direction x, y and z, and a uniform time step $\delta t = 0.1$. Then δt is successively halved ($\frac{\delta t}{2^i}$, $i = 1, 2, ...$), up to $\delta t = 0.0015625$. The final time was set to be 1 for all simulations.

The order of convergence is estimated by dividing the errors above, computed for two sets of parameters. Dividing the natural logarithm of the result by the natural logarithm of the refinement ratio yields an approximation of the convergence order, in the other term, if we denote by $\varepsilon_{\delta t}$ the error obtained using time step δt, the convergence rate will be approximatively equal to

$$\sigma_{\frac{\delta t}{2}} = \frac{\log \frac{\varepsilon_{\delta t}}{\varepsilon_{\frac{\delta t}{2}}}}{2}.$$

Here again, Table 3 shows that the convergence rate is 1, whereas the theoretical result above shows that the rate order is equal to $\frac{1}{2}$. In Table 4, we present the errors $E_u = \|u - u_{N\tau}\|_{L^2(\Omega)}$ and $E_q = \|\mathbf{q} - \mathbf{q}_{N\tau}\|_{L^2(\Omega)}$ at the final time $T = 1$.

τ	E_u	$E_{1,u}$	$E_{\partial_t u}$
0.1	0.219389	0.797017	1.8393
0.05	0.104713	0.487205	0.536101
0.0125	0.023401	0.120181	0.112778
0.00625	0.0113393	0.0588985	0.0544602
0.003125	0.00558312	0.0291962	0.0267956
0.0015625	0.00277492	0.0146111	0.0133426

Table 3: The errors for decreasing time steps.

N	E_u	$E_{1,u}$
5	0.00030507	0.00690947
7	7.99485e-05	0.00311564
9	2.44149e-05	0.00135193
11	8.12011e-06	0.00056492
13	2.65983e-06	0.000229697
15	8.86705e-07	9.14802e-05
17	2.95653e-07	3.58423e-05
19	9.89926e-08	1.38593e-05

Table 4: The errors for increasing polynomial degree N.

Finally, in Figures 12 and 13, we present the errors between exact solution $(u, \mathbf{q})$ and spectral approximation $(u_{N\tau}, \mathbf{q}_{N\tau})$ in each sub-domains Ω_k, where all errors are uniform. Note that, in these figures, the solutions are plotted in spectral grids when $N = 20$ in all directions and in each Ω_k.

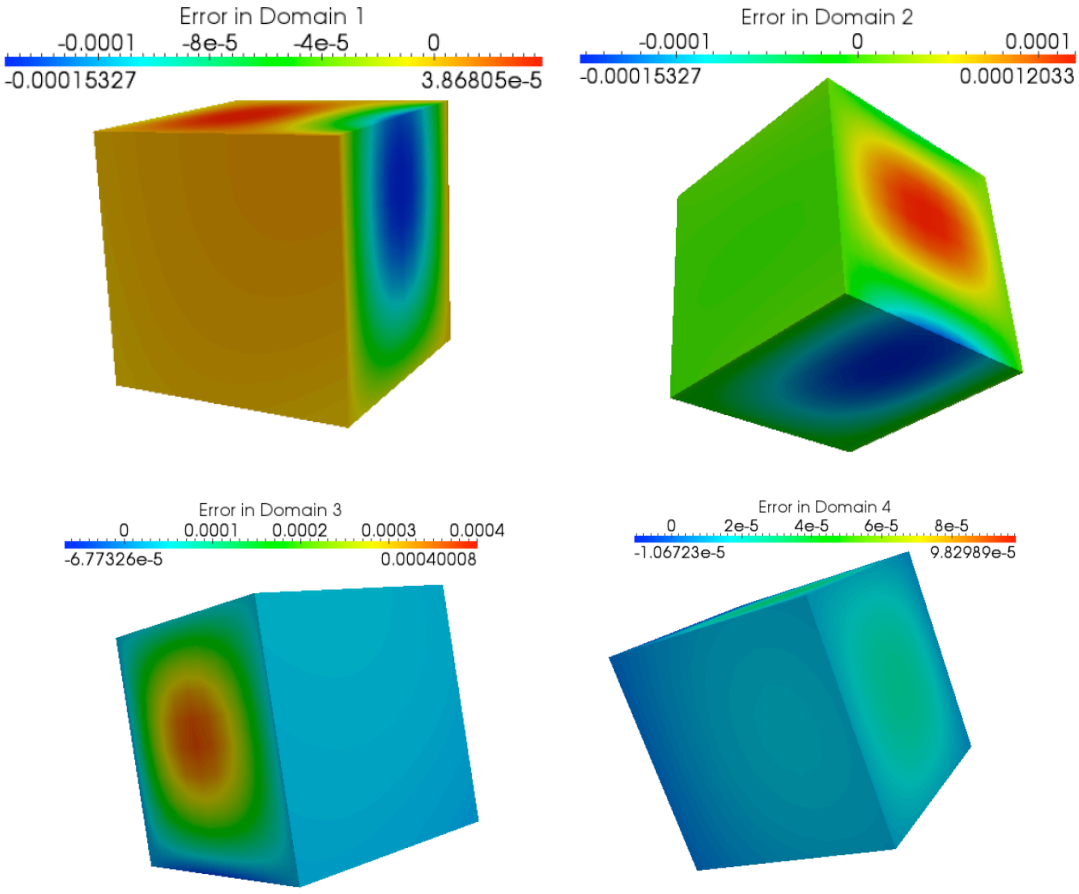

Figure 12: Errors on u at $T = 1$ with $\delta t = 0.0015625$.

All these results are in good coherence with the estimates (8.1). They confirm the efficiency of the spectral element method for solving the Richards equation in complex geometries.

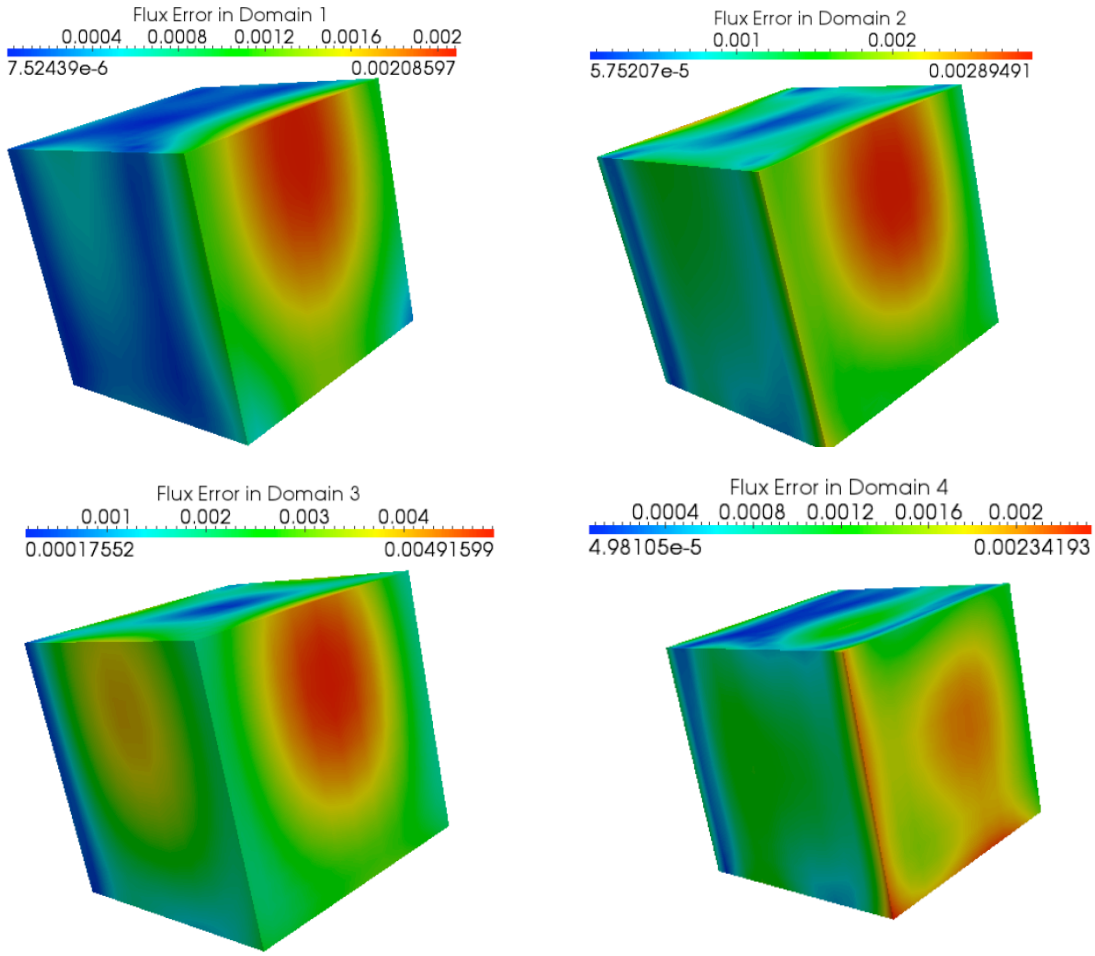

Figure 13: Errors on $\mathbf{q}$ at $T = 1$ with $\delta t = 0.0015625$.

Acknowledgments

We are deeply grateful to Linda El Alaoui who revealed us the existence of Richards equation and showed her high interest for our work.

References

[1] H.W. Alt, S. Luckhaus — Quasilinear elliptic-parabolic differential equations, *Math. Z.* **183** (1983), 311–341.

[2] M. Azaïez, F. Ben Belgacem, M. Grundmann, H. Khallouf — Staggered grids hybrid- dual spectral element method for second order elliptic problems. Application to high-order time

splitting for Navier-Stokes equations, *Compu. Meth. Appl. Mech. and Engrg.* **166** (1998), 183–199.

[3] J.-M. Bernard — Density results in Sobolev spaces whose elements vanish on a part of the boundary, *Chinese Ann. Math. Ser. B,* **32** (2011), 823–846.

[4] C. Bernardi, L. El Alaoui, Z. Mghazli — A posteriori analysis of a space and time discretization of a nonlinear model for the flow in partially saturated porous media, *IMA J. Numer. Anal.* **34** (2014), 1002–1036.

[5] C. Bernardi, Y. Maday — *Spectral Methods*, in the *Handbook of Numerical Analysis* **V**, P.G. Ciarlet & J.-L. Lions eds., North-Holland (1997), 209–485.

[6] C. Bernardi, Y. Maday, F. Rapetti — *Discrétisations variationnelles de problèmes aux limites elliptiques*, Collection "Mathématiques et Applications" **45**, Springer-Verlag (2004).

[7] H. Brezis, P. Mironescu — Gagliardo–Nirenberg, composition and products in fractional Sobolev spaces, *J. Evol. Equ.* **1** (2001), 387–404.

[8] F. Brezzi, J. Rappaz, P.-A. Raviart — Finite dimensional approximation of nonlinear problems, Part I: Branches of nonsingular solutions, *Numer. Math.* **36** (1980), 1–25.

[9] S. Del Pino and O. Pironneau — A fictitious domain based on general pde's solvers, *Proc. ECCOMAS 2001*, Swansea, K. Morgan Ed., Wiley (2002).

[10] R. Eymard, M. Gutnic, D. Hilhorst — The finite volume method for Richards equation, *Computational Geosciences* **3** (1999), 259–294.

[11] M. Gabbouhy — Analyse mathématique et simulation numérique des phénomènes d'écoulement et de transport en milieux poreux non saturés. Application à la région du Gharb, Ph.D. Thesis, Université Ibn Tofail, Kénitra, Maroc (2000).

[12] S.M.F. Garcia — Improved error estimates for mixed finite-element approximations for nonlinear parabolic equations: the continuous-time case, *Numer. Methods Partial Differential Equations* **10** (1994), 129–147.

[13] S.M.F. Garcia — Improved error estimates for mixed finite-element approximations for nonlinear parabolic equations: the discrete-time case, *Numer. Methods Partial Differential Equations* **10** (1994), 149–169.

[14] V. Girault, P.-A. Raviart — *Finite Element Methods for Navier–Stokes Equations, Theory and Algorithms*, Springer–Verlag (1986).

[15] J.-L. Lions — *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod & Gauthier-Villars (1969).

[16] M J.-L. Lions, E. Magenes — *Problèmes aux limites non homogènes et applications*, Vol. I, Dunod, Paris (1968).

[17] Y. Maday, E.M. Rønquist — Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries, *Comput. Methods in Applied Mech. and Engrg.* **80** (1990), 91–115.

[18] J.-C. Nédélec — Mixed finite elements in $\mathbb{R}^3$, *Numer. Math.* **35** (1980), 315–341.

[19] I.S. Pop, F. Radu, P. Knabner — Mixed finite elemets for the Richards' equations: linearization procedure, *J. Comput. Appl. Math.* **168** (2004), 365–373.

[20] F. Radu, I.S. Pop, P. Knabner — Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation, *SIAM J. Numer. Anal.* **42** (2004), 1452–1478.

[21] K.R. Rajagopal — On a hierarchy of approximate models for flows of incompressible fluids through porous solid, *Math. Models Methods Appl. Sci.* **17** (2007), 215–252.

[22] L.A. Richards — Capillary conduction of liquids through porous mediums, *Physics* **1** (1931),
318–333.

[23] Y. Saad and M. H. Schultz — GMRES : a generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Statist. Comput.* **7(3)** (1986), 856–869.

[24] M. Schatzman — *Analyse numérique*, InterEditions, Paris (1991).

[25] E. Schneid, P. Knabner, F. Radu — A priori error estimates for a mixed finite element discretization of the Richards' equation, *Numer. Math.* **98** (2004), 353–370.

[26] M. Slodicka — A robust and efficient linearization scheme for doubly nonlinear degenerate parabolic problems arising in flow in porous media, *SIAM J. Sci. Comput.* **23** (2002), 1593–1614.

[27] P. Sochala, A. Ern, S. Piperno — Mass conservative BDF-discontinuous Galerkin/explicit finite volume schemes for coupling subsurface and overland flows, *Comput. Methods Appl. Mech. Engrg.* **198** (2009), 2122–2136.

[28] P. Sochala, A. Ern, S. Piperno — Numerical methods for subsurface flows and coupling with surface runoff, in preparation.

[29] C.S. Woodward, C.N. Dawson — Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media, *SIAM J. Numer. Anal.* **37** (2000), 701–724.

[30] D. Yakoubi — Analyse et mise en øuvre de nouveaux algorithmes en méthodes spectrales. Ph.D. Thesis, Université Pierre et Marie Curie, Paris, France (2007).